Use of Collateral Information to Improve Landsat Classification Accuracies

NASA Grant NSG-2377

Semiannual Progress Report

Alan H. Strahler
and
John E. Estes
Co-Principal Investigators
Geography Remote Sensing Unit

University of California
Santa Barbara, CA
93106

Technical Monitor: Dr. David L. Peterson, NASA-Ames

UNIVERSITY OF CALIFORNIA, SANTA BARBARA

BERKELEY • DAVIS • IRVINE • LOS ANGELES • RIVERSIDE • SAN DIEGO • SAN FRANCISCO SANTA BARBARA • SANTA CRUZ

SANTA BARBARA, CALIFORNIA 93106

June 17, 1980

Dr. David L. Peterson
NASA-Ames Research Center
Mail Stop 242-4
Moffett Field, CA 94035

Dear Dave:

The purpose of this letter is to provide a semi-annual report of our activities under NASA Grant NSG-2377. The period covered by the report will be from October 1, 1979 to March 30, 1980, the six months immediately proceeding our second year renewal on April 1. The semi-annual report is submitted in fulfillment of our responsibilities as described in "NASA Provisions for Research Grants." My letter of December 10, 1979, serving formally as a report of activities during the July 1 to September 30, 1979 interval, also discusses some of the activities which have occurred in the October to December time period. Although I will refer to these activities in this report, I will emphasize activities occurring after the first of the year.

Publications. I am very pleased to announce that during this period three manuscripts have been placed in the publication process. The first of these, entitled "The Use of Prior Probabilities in Maximum Likelihood Classification of Remotely Sensed Data," was enclosed with my letter of December 10. This manuscript has, within the past few weeks, been accepted for publication in Remote Sensing of Environment. Since the manuscript was accepted without revision, it should appear essentially the same as the copy which you possess.

The remaining two manuscripts are in symposium proceedings. The first of these is "Incorporating Collateral Data in Landsat Classification and Modeling Procedures," which will appear in the Proceedings of the Fourteenth International Symposium on Remote Sensing of the Environment. The second is "A Logit Classifier for Multi-Image Data," which will appear in the Proceedings IEEE Workshop on Picture Data Description. I enclose copies as specified in the grant provisions.

Travel. In March, 1980, my research associate, Mr. Curtis Woodcock, and I attended an informal meeting of California researchers involved in forestry applications of remote sensing at California Polytechnic Institute and State University, San Luis Obispo. At your invitation, we briefed the group on some of our current forestry work as well as the implications of our research concerning the incorporation of collateral data into Landsat processing systems. We enjoyed this constructive opportunity to share research ideas with investigators working on similar applications.

Grant travel funds were also used to support a side trip from Denver, Colorado to Fort Collins in December, 1979, to allow me to confer with Dr. James A. Smith, of the School of Forestry, concerning spectral modeling and possible joint research in exploring the role of collateral information in the remote sensing process.

Although not during the time period covered by this report, I also attended the Fourteenth International Symposium on Remote Sensing of the Environment, held in San Jose, Costa Rica, in April, 1980. At this meeting, I had many productive exchanges with remote sensing researchers, and was allowed the opportunity to present the results of our first year's work in this grant. I should also mention that, as a result of discussions at that meeting, I have been invited to present a seminar concerning the use of collateral information in remote sensing at the Canada Centre for Remote Sensing (CCRS) in late June. Travel and per diem expenses for this trip will be reimbursed by CCRS.

Technical Progress. During this reporting period, we made substantial progress in three research topics. The first of these is the survey of quantitative techniques available for merging categorical and continuous variables, and the assessment of the utility of these techniques for remote sensing applications. This survey was completed, and the results are summarized in the first part of the San Jose preprint (attached). The second area of progress concerns the further application of prior probabilities in the context of a time-dependent land use classification system. Again, the results are summarized in the San Jose preprint.

Perhaps some of our most exciting work concerned the development and application of the logit model to remote sensing. This model does not rely on multivariate normal assumptions, and allows probability of class membership to be predicted as a function of any array of independent variables, either categorical or continuous in nature. Further, the logit model may be formulated in a linear or curvilinear fashion, depending on the application. Our research in this area is summarized in the attached preprint, "A Logit Classifier for Multi-Image Data."

Software. During the reporting period, we have also developed two new software items. The first of these is a program which fits a logit model to a set of calibration data which are input to the program. Although existing programs are available for logit modeling, none of these were suitable for the data which were available. The second software item was a new VICAR program, PROBMAPS, which accepts images of independent variables, input in MSS format, as well as the calibration parameters for the logit model, and produces a set of output images recording the probability of membership of each pixel for each class as predicted by the logit model. Both of these programs are quite specific to the logit model application, but both have been coded in universal FORTRAN so that they can be compiled at other institutions. We will be happy to make this software available to you at NASA-Ames or to any other users as well.

Future Plans. Our plans for the coming research period have not been completely formalized at this time. I hope to be able to pursue most or all of the following research topics:

1. modeling universal soil loss for a pixel based on landform, land use, rainfall, and topographic parameters;
2. improved rainfall runoff modeling on a pixel-by-pixel basis for use in hydrologic modeling;
3. further exploration of the logit model including its ability to simulate various nonnormal distributions and perform as a classifier of remotely sensed data; and,
4. continue the exploration of time-sequential classification of land use in Ventura County.

This concludes my semi-annual report.

Sincerely yours,

Alan H. Strahler
Principal Investigator

A LOGIT CLASSIFIER FOR MULTI-IMAGE DATA

Alan H. Strahler and Paul F. Maynard

Preprint: Proc. IEEE Workshop on Picture Data Description, 8/80, Asilomar, CA.

Department of Geography
University of California
Santa Barbara, California 93106

A new classifier for multi-image databases uses maximum likelihood estimation of parameters fitting a logit model to training data. A logit is the natural logarithm of a probability ratio: e.g., $\ln (P_1/P_0)$. As an example, a linear logit classification model for a simple two-class case based on four Landsat channels is:

$$\ln(\frac{P_a}{P_b}) = \beta_0+\beta_1(LS4)+\beta_2(LS5)+\beta_3(LS6)+\beta_4(LS7)$$

where P_a and P_b are probabilities that the pixel belongs to class a and b respectively, $\beta_0...\beta_4$ are calibration constants, and LS4...LS7 and the four Landsat channels for bands 4 through 7.

Compared with usual Bayesian maximum likelihood classification, the logit classifier has certain distinct advantages. It is nonparametric, in that multivariate normality is not assumed. The model may be specified in linear or curvilinear forms as appropriate. Further, the model can incorporate categorical information in the form of dummy variables, and can therefore be used to merge continuously measured image data with categorical collateral data in a single classification step.

Introduction

The Bayes maximum likelihood classifier (MLC) is the most commonly used decision rule for discrete classification with Landsat derived data. Such a classifier uses the Bayes decision rule to assign pixels to the class with highest probability, given the observed vector of spectral measurements and the prior distribution of classes. To find this probability, the Bayes MLC requires an estimate of the conditional probability of occurrence of the observed vector of MSS data given that it is associated with a specified class. This estimate has traditionally been obtained by assuming that the observed measurement vectors were Gaussian, or normal; therefore, the best estimate is a measure of the probability density value of the multivariate normal distribution for the class evaluated at the observation.

With four channels of Landsat spectral data, the Bayes maximum likelihood classifier has achieved good classification accuracies in many cases. Classification accuracies have been further boosted by the inclusion of collateral data as additional logical channels or as indexes to sets of prior probabilities in the case of categorical collateral variables.[1] Further increases in classification accuracy will undoubtedly result from more optimal spectral channels and improved techniques of incorporating collateral data.

However, it is very likely that the spectral reflectances from some classes in many applications are not normally distributed. In this situation, even with the ideal spectral "windows" and precise and relevant collateral data, the Bayes MLC will reach an asymptotic accuracy limit that will be less than optimal. In fact, for classes that deviate from normality, classification accuracies could be significantly less than optimal.

Furthermore, each of the two previously mentioned techniques of increasing accuracy by input of collateral data to the Bayes classifier has critical limitations. The Bayes MLC can only accept measurement variables which are continuously distributed, and this requires continuous measurement (or at least discrete measurement with a large number of discrete steps). Consequently, the collateral channel in the direct input approach can utilize only data which are measured on the interval or ratio scale. This requirement eliminates many potentially useful databases, such as soils maps, geologic maps, political boundaries, census tracts, etc. And when collateral information is incorporated through the mechanism of prior probabilities, the calculation of the prior probabilities usually requires a sophisticated sampling design.

One solution to these problems is to use a statistical technique that predicts probabilities of categorical membership with no distribution assumptions. One such technique, called the logit regression model, has been widely used in the social sciences for the last twenty years. The logit regression model generates predicted probabilities that all sum to one for a specified suite of classes, and the classification can be awarded to the category with the highest predicted probability. Further, predictor variables may be continuous or categorical; the model may be specified in linear or curvilinear forms; and the assumption of multivariate normality is not required.

This paper has the following components:

1. a review of the mathematics of the Bayes MLC;
2. an examination of the sources of classification error due to non-normal distributions;

3. an examination of the problems encountered with utilizing collateral data;
4. development and exposition of the logit regression model, with an emphasis on its ability to act as a nonparametric classifier;
5. a description of planned research which will compare the classification accuracies of the logit regression model and the Bayes MLC for a land use/land cover classification example; and
6. presentation of an example of the use of linear logit model in a remote sensing application.

The Bayes Maximum Likelihood Classifier

Background

In the past ten years, maximum likelihood classification has found wide application in the field of remote sensing. Based on multivariate normal distribution theory, the maximum likelihood classification algorithm has been in use for applications in the social sciences since the late 1940's. Providing a probabilistic method for recognizing similarities between individual measurements and predefined standards, the algorithm found increasing use in the field of pattern recognition in the following decades.[2,3,4] In remote sensing development of multi-spectral digital images of land areas from aircraft or spacecraft provided the opportunity to use the maximum likelihood criterion in producing thematic classification maps of large areas for such purposes as land use/land cover determination and natural cultivated land inventory.[5,6]

Derivation of the Bayes MLC

In order to understand the difference between a classification awarded on the basis of logit-generated predicted probabilities and posterior probabilities derived from the Bayes MLC, it will be helpful to briefly review the mathematics of the Bayes maximum likelihood decision rule. In the multivariate remote sensing application, it is assumed that each observation X (pixel) consists of a set of measurements on p variables (channels). Through the selection of training sites, a set of observations which correspond to a class is identified -- that is, a set of similar objects characterized by a vector of means on measurement variables and a variance-covariance matrix describing the interrelationships among the measurement variables which are characteristic of the class. Although the parametric mean vector and dispersion matrix for the class remain unknown, they are estimated by the sample means and dispersion matrix associated with the object sample.

Multivariate normal statistical theory describes the conditional probability that an observation X will occur, given that it belongs to a class i, as the following function:

$$P(X|\omega_i) = (2\pi)^{-\frac{1}{2}p}|\Sigma_i|^{-\frac{1}{2}}e^{-\frac{1}{2}(X-\mu_i)'\Sigma_i^{-1}(X-\mu_i)} \quad (1)$$

(Please refer to Table 1 for a description of the mathematical symbols). As applied in a maximum likelihood decision rule, expression (1) allows the calculation of the conditional probability that an observation is a member of each of k classes. However, the actual probability desired is the posterior probability $P(\omega_i|X)$; it can be shown[4] that:

$$P(\omega_i|X) = \frac{p_i(X)\cdot P(\omega_i)}{\sum_{i=1}^{k} p_i(X)\cdot P(\omega_i)} \quad (2)$$

This expression leads to the decision rule:

Choose i which minimizes

$$\ln|D_i| + (X-m_i)'D^{-1}(X-m_i) - 2\ln P(\omega_i), \quad (3)$$

In usual practice the prior probabilities $P(\omega_i)$ are assumed equal, or $P(\omega_i)=1/R$ where R is the number of classes. In this case, the last term in expression (3) is constant over all R classes, and need not be considered in the decision rule. This equal priors decision rule is used in the currently distributed versions of LARSYS and VICAR, two image processing systems authored respectively by the Laboratory for Applications of Remote Sensing at Purdue University and the Jet Propulsion Laboratory of California Institute of Technology at Pasadena.

Classification Accuracy and the Assumption of Normality

In a typical supervised classification, training sites are selected by the analyst to typify each class. Histograms of spectral values for classes are inspected for multivariate normality, and when a class actually consists of several distinctive signatures, training sites are reaggregated into subclasses, each of which is approximately multivariate normal. In this way, a set of multivariate normal dispersion patterns are defined for the desired classes. It is important to realize that such dispersions, because they are selected to be as "pure" as possible, are probably underdispersed with respect to the true information class. This effect produces a difficulty in the classification of mixed pixels. Since the MLC model does not provide for mixed pixels, the implicit assumption is that mixed pixels are to be classified according to the most probable signature match; the components of the signature which are reflected from the less important classes contained within the pixel are thus regarded as random noise. The mixed pixel, then, is typically classified by comparing probability densities within the tails of overlapping multivariate normal distributions. The accurate classification of mixed pixels under MLC thus requires a good fit of the tails to multivariate normality; however, it is obvious that the selection of training sites for purity will of necessity produce a poor fit in the tails. And, mixed pixels will constitute a large portion of the scene -- up to forty percent in some agricultural applications (F. Hall, personal communications).

This reasoning naturally leads to the consideration of nonparametric classifiers. Brooner et al.[7] compared the Bayes MLC to a classifier which used $P(X|\omega_i)$ as directly estimated by a sampling procedure, and reported a four percent increase in classification accuracy over MLC. However, direct estimates of $P(X|\omega_i)$ require more data as well as

dealing with p-dimensional table storage problem. The logit model, discussed in the following pages, provides an alternative which should require fewer data to calibrate and can, by virtue of its curvilinear modeling, approximate the real distribution without assuming multivariate normality.

Problems with Incorporating Prior Probabilities

As shown earlier, the Bayes MLC can easily be modified to take into account prior probabilities which describe how likely a class is to occur in the population of observations as a whole. The prior probability itself is simply an estimate of the proportion of pixels which will fall into a particular class. These prior probabilities are sometimes termed weights, since the modified classification rule will tend to weight more heavily those classes with higher prior probabilities. Strahler[1] showed via simplified numerical examples how these different weights can affect the decision of the Bayes MLC. As the prior probability $P(\omega_i)$ in expression (3) becomes large and approaches 1, its logarithm will go to zero and the classification decision will effectively be made with expression (1).

However, since this possibility and all others must sum to one, the prior probabilities of the remaining classes will be small numbers, thus increasing the value of the expression. Since the classifications is ararded to the class with the smallest value, the effect will be to force classification into the class with high probability. Therefore, the more extreme are the values of the prior probabilities, the less important are the actual observation values X.

Strahler[1] has demonstrated how prior probabilities can be used as a mechanism to incorporate collateral data in categorical from into the Bayes MLC. His mechanism uses a set of prior probabilities estimated for each collateral category by an external sampling procedure, with the classification algorithm accessing the appropriate set of probabilities contingent upon the collateral category of the pixel. In this fashion, categorical collateral information is merged with multivariate normal information concerning the spectral signatures. Although this approach was proven effective for a forestry application,[1] estimations of the sets of prior probabilities may require considerable data collection, depending on the number of classes and collateral categories. In addition, multivariate normality of signatures is assumed, and the comments of preceding paragraphs apply. In contrast, use of the logit classification model allows categorical and continuous information to be mixed freely through the mechanism of dummy variables. And, again, the model can be fitted in a linear or curvilinear fashion as desired, avoiding the assumption of multivariate normality. Thus, the logit model offers a more natural, straightforward way of incorporating categorical variables into the classification procedure.

The Logit Regression Model

Linear Modeling of Probabilities

The most commonly used predictive multivariate statistical technique is probably ordinary least squares (OLS) regression. The prediction, or estimated value of the dependent variable, is a function of the vector of estimated betas ($\hat{\beta}$) in combination with the vector of observed independent variables (X). The betas are estimated in such a way that the variance about the least squares regression line is minimized.

When used to model probabilities, OLS regression has one major drawback: although probabilities are constrained to lie within the range of 0 to 1, the predictions generated from such a model are unbounded and may take values from minus infinity to plus infinity. Thus, the predictions may lie outside the meaningful range of probability. Further, the probability of each class must be modeled separately, and there is no constraint to ensure that all probabilities must sum to one.

One solution to the bounding problem is to specify that

$$0 \leq P_i \leq 1$$

(where P_i is the probability of observing a specified class or category of the dependent variable). In the case of ordinary least squares regression model,

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2}.$$

The simplest way to satisfy this condition is to impose the following arbitrary definition of P_i:

1. P_i is equal to 0 if Y_i is less than 0;
2. P_i is equal to Y_i if Y_i is equal to or between 0 and 1.
3. P_i is equal to 1 if Y_i is greater than 1:

and use straightforward ordinary least squares estimation of the regression parameters. This solution is often referred to as the linear probability model. Unfortunately, although it appears to be a simple solution to the predicted probabilities problem, the model has a number of serious limitations which are discussed by Domencich and McFadden.[2] Again, the probabilities are not constrained to sum to one.

The Logit Model

The simplest, yet most statistically sound, solution to the probability problem (within a regression framework) is the logit transformation. In this transformation, the ratio between the probability that an observation or pixel i belongs to a class P_1 and the probability that it does not belong to P_1 is expressed as a logistic function:

$$\frac{P_{i,1}}{1-P_{i,1}} = e^{\beta_1 X} \qquad (4)$$

where β is a vector of parameters and X is a vector of observations on independent variables. Taking the natural logarithm of this expression,

$$\ln\left(\frac{P_{i,1}}{1-P_{i,1}}\right) = \beta_1 X \quad . \tag{5}$$

The left-hand quantity is referred to as a logit. Note that when βX is zero, the ratio will be 1, indicating equal probability. As βX varies positively or negatively, the ratio will shift accordingly.

The ratio, under the constraint that the numerator and denominator must sum to one, determines the two probabilities uniquely. Expression (4) can be solved explicitly for P_1:

$$P_{i,1} = \frac{e^{\beta_1 X}}{1+e^{\beta_1 X}} \quad . \tag{6}$$

And, if P_2 is defined as $1-P_1$, it is easy to show that

$$P_{i,2} = 1-P_{i,2} = \frac{1}{1+e^{\beta_1 X}} \quad . \tag{7}$$

Although expressions (4) and (5) show the product βX as a linear function, the X vector may contain powers and cross products in the case of a curvilinear model. An example is (elemental notation):

$$\ln\left(\frac{P_i}{1-P_i}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{1,1} X_1^2 + \beta_{2,2} X_2^2 + \beta_{1,2} X_1 X_2 \tag{8}$$

for the bivariate case.

Unlike the conventional regression models, the logistic and linear logit regression models require either a weighted least squares (WLS) procedure or a maximum likelihood procedure to estimate the calibration parameters (betas). The choice between the two methods depends upon whether or not the sample under investigation includes repeated observations for each combination of values of the explanatory variables. If so, WLS is appropriate; however, remote sensing applications rarely have repeated observations. Consequently, the method of maximum likelihood is the preferred method. A number of authors present the details of this method, which is discussed briefly in a following section.[9,10,11,12]

Maximum likelihood estimation of logit model parameters has many other attractive features. Provided that the sample data are not multicollinear, a unique maximum likelihood estimator can be obtained even in relatively small samples. Also, the mathematical properties of the likelihood function allow for efficient computer programs to produce the parameter estimates. These estimates are consistent and are the best possible estimates in very large samples. The disadvantages of the procedure are that it involves numerical optimization and therefore more computation, and that it requires more calibration data than MLC because of the larger number of parameters which need to be estimated.

Logit Example

The following is an example that uses continuous and categorical explanatory variables to estimate

$$\ln\left(\frac{P_{i,1}}{1-P_{i,1}}\right) = \beta_0 + \beta_1 X_{i1} + D_1 \beta_2 \quad , \tag{9}$$

where $P_{i,1}/(1-P_{i,1})$ is the ratio of the probability that pixel i is not of class 1, X_{i1} refers to a continuously measured variable on pixel i (MSS or continuous collateral data), and D_1 is a dummy variable that is equal to one if a categorical variable is true at pixel i and zero if it is not. Given the logit ratio, simple algebra will extract the value of $P_{i,1}$. It is straightforward to add more continuous and categorical explanatory variables.

Estimating the Regression Parameters

In order to calculate the logit ratio in the preceding formula, it is necessary to obtain estimates of the regression parameters. The first step is to specify the model in terms of a likelihood function. If the training observations are thought of as independent trials, then the likelihood of the outcome of these trials (for the two-class case described above) is:

$$L = \prod_{i=1}^{n_1} P_{i,1} \prod_{i=n_1+1}^{n} (1-P_{i,1}) \tag{10}$$

where observations $i=1$ to n_1 are those in which the observed dependent variable was a member of class 1 and $i=n_1+1$ to n are the observations in which the dependent variable was not a member of class 1. Substituting from the definitions of $P_{i,1}$ and $1-P_{i,1}$ in expressions (6) and (7), the result is

$$L = \prod_{i=1}^{n_1} \frac{e^{\hat{\beta} X_i}}{1+e^{\hat{\beta} X_i}} \prod_{i=n_1+1}^{n} \frac{1}{1+e^{\hat{\beta} X_i}}. \tag{11}$$

As specified, the likelihood depends upon a set of unknown parameters, the betas. These parameters are estimated by choosing those values which maximize the preceding likelihood formula for the given set of training data. Rather than maximize the likelihood itself, it is computationally simpler to maximize the logarithm of the likelihood, or

$$\ln L = \sum_{i=1}^{n_1} \hat{\beta} X_i - \sum_{i=1}^{n} \ln(1+e^{\hat{\beta} X_i}). \tag{12}$$

To maximize this expression it is not possible to set the partial derivatives of ln (L) with respect to the betas to zero, and solve simultaneously for the betas in a direct fashion. Instead, the solution must be obtained by iteratively recalculating ln L for successive estimates of betas until the partial derivatives converge upon zero.

There are several mathematical techniques for iteratively converging the first partial derivatives to zero. One well-known technique is the Newton-Raphson Method, which calculates deltas (amount of change) for the betas by forming the matrix of second partial derivatives, inverting it, and postmultiplying it by the vector of the first partial derivatives. That is,

$$\underline{\delta} = \underline{\Sigma}^{-1}\underline{g},$$

where $\underline{\delta}$ is the vector of calculated deltas, $\underline{\Sigma}^{-1}$ is the inverted second partial derivative matrix, and $\underline{g}$ is the vector of first partial derivatives. The deltas are subtracted from the betas, and the second partial derivative matrix and first partial derivatives are recalculated with the new betas. The process continues until the vector of first partial derivatives has converged upon zero, at which point the most likely vector of betas has been identified.

Given these maximum likelihood estimates of the betas, the last step in a logit model classification sequence is to use expressions (6) and (7) to calculate the vector of probabilities for each pixel and award the classification to the category with the largest predicted probability.

Polychotomous Logit Regression

The preceding example, although conceptually straightforward, is not applicable when there are more than two categories to be predicted. The dichotomous logit can be easily extended to the polychotomous logit. Now, instead of two categories of interest, there are R possible categories. The model now becomes:

$$P_{i,r} = \frac{e^{\beta_r X_i}}{\sum_{r=1}^{R} e^{\beta_r X_i}}, \qquad (13)$$

where $P_{i,r}$ is the probability that pixel i belongs to the rth category.

Because of the constraint that the probabilities must sum to one, only $R-1$ sets of betas and probability ratios need to be determined. Introducing this constraint on the Rth class, it is easy to show algebraically that

$$P_{i,R} = 1 - \sum_{r=1}^{R-1} P_{i,r} = \frac{1}{1+\sum_{r=1}^{R-1} e^{\beta_r X_i}}. \qquad (14)$$

This constraint can also be introduced by taking $\beta_R = 0$, which produces an identical expression from substitution into expression (13).

For estimates of the betas, the maximum likelihood estimation procedure, described above for the two-class case, is generalized to the R-class case in a straightforward manner.

Proportion Estimation

Although the discussion above has stressed the use of the logit model for classification, it may also be used for proportion estimation. Nelepka et al.[13] and Woodcock et al.[14] have both discussed this problem using underlying assumptions of multivariate normality. The logit model provides an alternative which does not assume multivariate normality and estimates proportions directly. As in classification, either linear or curvilinear models may be selected, and categorical variables may be readily utilized as well. The difference between application of the logit model as a classifier and as a proportion estimator lies in the nature of the calibration data. For the classifier, training observations of $\underline{x}$ vectors (for pixel i) are each individually labeled with a single class; in the case of proportion estimation, each observation contains the observed proportions of classes and the associated measurement vectors. These proportions constitute weights, and it is easy to show that the likelihood function becomes (as in the two-class case):

$$\ln(L) = w\sum_{i=1}^{n} p_{i,2}\beta_1 X_{i,1} + w\sum_{i=1}^{n} p_{i,2}\beta_2 X_{i,2} - w\sum_{i=1}^{n} \ln(1 + e^{\beta_1 X_{i,1} + \beta_2 X_{i,2}}),$$

where $p_{i,k}$ is the proportion for the ith observation for class k, and w is the observation weight (which, as a constant, is eliminated in differentiation of the fraction).

Klamath National Forest Test Site Location

Figure 1. Index map showing location of area modeled in Klamath National Forest.

Application Example

At the present time, the logit-based classifier has not been tested, although a logit model has been used in a forestry remote sensing problem of proportion estimation. This use is summarized in the paragraphs below. For this application, a linear logit model was devised and fitted to forest species compositional data for northern California, predicting the proportion of timber volume for each of five coniferous tree species at each pixel based on registered terrain data quantifying elevation, slope, and aspect. This research utilizes the Video Image Communication and Retrieval (VICAR) system and the Image Based Information System (IBIS) resident at the University of California, Santa Barbara. VICAR/IBIS, developed at the Jet Propulsion Laboratory (JPL) at Pasadena, California, is a job control language which permits the sequential linking and execution of a vast array of Fortran and Assembler routines in a batch environment. In addition to extensive usage of existing VICAR/IBIS routines, new VICAR and non-VICAR software were developed and/or modified as required for this application.

Logit modeling of species proportions used data derived from the Klamath National Forest, located in northern California (Figure 1). Ranging in relief from 500 to 8,000 feet, the Forest includes 2,600 square miles of rugged terrain in the Siskiyou, Scott Bar, and Salmon Mountains. Little of the area is developed beyond management for timber yield, livestock production, and recreation. A wide variety of distinctive vegetative types is present in the area. Forest vegetation includes such coniferous species as noble, red, white, and douglas firs, ponderosa pine, and incense cedar, as well as several oaks, and typical species of chaparral. Thus, the topographic and vegetational characteristics of the area are well differentiated. Within the Klamath National Forest, a study area including most of the Goosenest Range was selected for logit modeling of species composition from terrain features. This area was chosen because calibration data and Landsat images were readily available for it.

Digital Terrain Model

The logit model for this forestry application requires preparation of digital terrain data. These data, obtained from the National Cartographic Information Center, in Reston, Virginia, are derived from processing of 1:250,000 contour maps, and include elevations at every point on a grid of approximately 65 m spacing. Although the data are comparable in scale to a Landsat image, the elevation values are quite generalized because they are produced from small scale contour maps by interpolation.

Slope angle and slope aspect channels can be produced using the elevation data of the registered terrain image. Although a number of slope and aspect generating algorithms are known, the simplest is the fitting of a least squares plane through each pixel and its four nearest neighbors and the calculation of the downslope angle and direction of the plane. Slope aspect was transformed from a coding of zero to 255 representing 0° to 359°to a cosine function shifted by 45°. This function, proposed by Hartung and Lloyd,[7] contrasts northeast-facing slopes, which present a favorable cool, moist growing environment, with hot, dry southwest-facing slopes. Although the function is defined ecologically, it also simulates Lambertian reflectance from a light source placed in the northeast, and thus the aspect image shown in Figure 2 gives the strong visual impression of relief.

Logit Model

The logit model fitted is

$$\ln\left(\frac{P_k}{P_{\neq k}}\right) = \hat{\beta}_{1,k} + \hat{\beta}_{2,k}E + \hat{\beta}_{3,k}A + \hat{\beta}_{4,k}S, \quad k = 1,5,$$

where P_k is the probability that a board-foot of timber volume will be drawn from one of five species k, $P_{\neq k}$ is the probability that the board-foot will not be drawn from species k, E is elevation (compressed to 0-255 range), A is aspect transformed as described above, S is slope angle, and $\hat{\beta}_{1,k},\ldots,\hat{\beta}_{4,k}$ are the estimated regression constants. Note that five equations, one for each species, actually comprise the model. The model was calibrated using 73 measurements of timber volume prepared by the U. S. Forest Service and located within two subregions of the Goosenest range. These samples are probably not representative of the entire area modeled, but serve for the demonstration purposes of this research. Each sample was located on 1:15,840 scale color air photos and transferred to Band 5 of a registered Landsat image to obtain the line and sample coordinates of the sample point. The coordinates were then used to extract elevation and aspect values for the sample from the registered elevation and aspect images. The coefficients for the model were fitted by a nonlinear optimization algorithm employing the Newton-Raphson method described above.

Given the constants produced by this procedure, the probability images were created using the new VICAR program "PROBMAPS." PROBMAPS, written specifically for this application, calculates the probability of species k for each pixel i using the following expression:

$$P_{i,k} = \frac{e^{\beta_k X_i}}{\sum_{j=1}^{5} e^{\beta_j X_i}}, \quad k = 1,5,$$

PROBMAPS then scales each probability so that the range 0=255 represents 0 to 1. PROBMAPS output images for this example are shown in Figure 2. Brightness values represent that probabilities occurrence for douglas fir, ponderosa pine, white fir, and red fir, with probabilities scales to range from black (0.) to white (1.0). The incense cedar image has been contrast stretched for display purposes, and presents a probability range of 0. to .3 from black to white. The probability images represent maximum likelihood estimates of species proportions; they appear reasonable in light of the known ecological preferences of the species, but their accuracies remain to be determined.

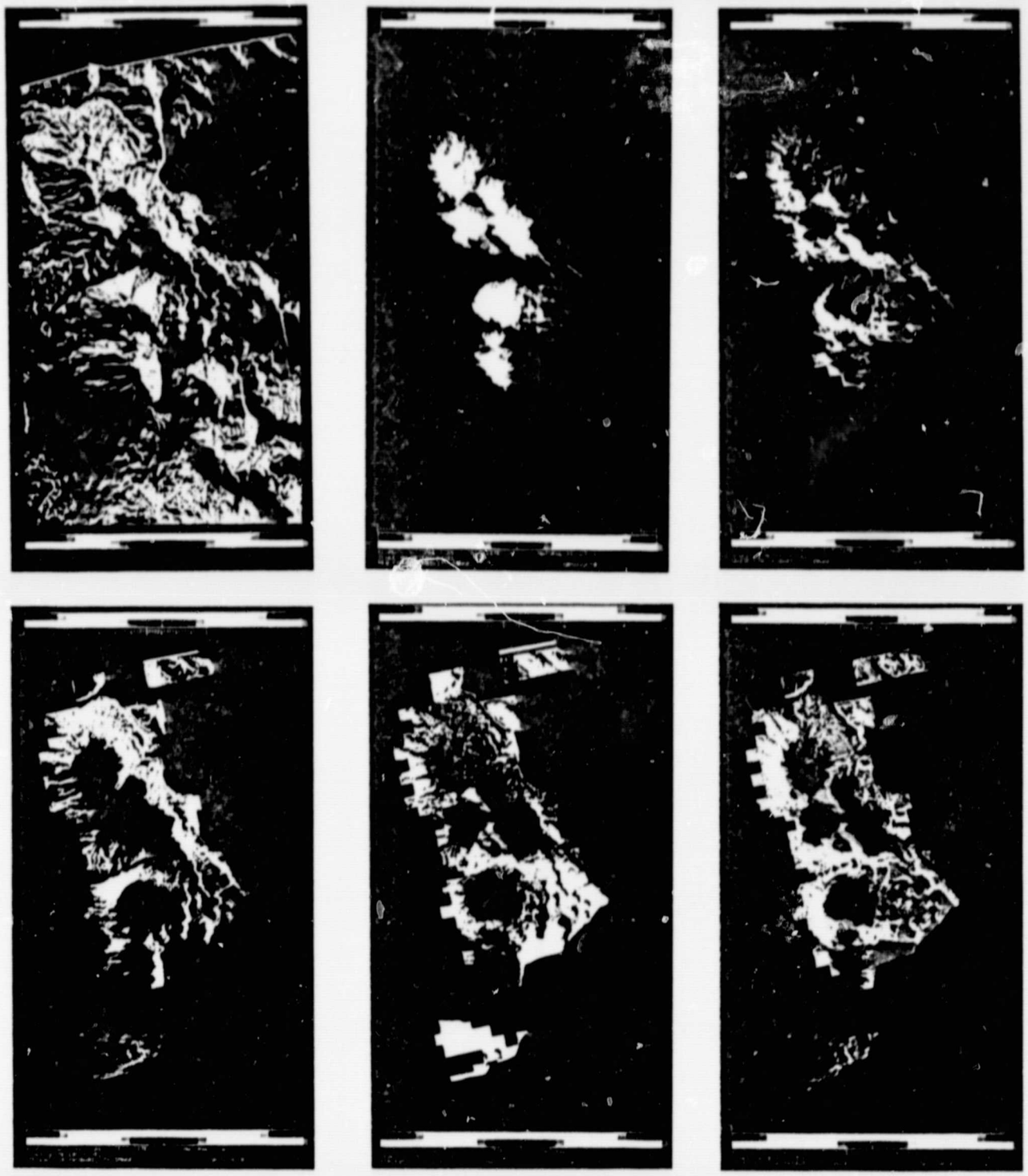

Figure 2. Clockwise from upper left: cosine function of slope aspect; probability images of coniferous forest species red fir; white fir; incense cedar; ponderosa pine; and douglas fir; for Goosenest test area within Klamath National Forest. For probability images, only area within Forest boundary is shown.

Future Work

Although the logit classifier appears to have some unique advantages [illegible] conventional maximum likelihood classification; further work will be necessary to prove its value for remote sensing applications. Topics to be investigated include:

1. Model Specification. For what shapes of non-normal distribution are linear models appropriate? Under what conditions are curvilinear models necessary? Could a stepwise procedure, analogous to polynomial curve fitting, be devised for model calibration? Since it is possible to obtain assymptotic estimates of the standard error of each beta, could the stepwise procedure drop individual terms from the model which are not significantly different from zero?
2. Accuracy. How well are probabilities predicted? Can a confidence limit be placed on the predicted probability? Monte Carlo methods may be helpful here. How does accuracy interact with distribution shape?
3. Further Applications. The logit model needs to be exercised on a real classification problem and compared with conventional MLC. Which is more accurate? Which consumes more computational resources? Do categorical variables present any special problems?

These questions and others will be the subject of future research in the application of the logit model to the remote sensing problem.

Table 1. Notation

Term	Definition
n	Number of measurement variables used to characterize each object or observation.
X	A p dimensional random vector.
X_i	Vector of measurements on n variables associated with the ith object or observation; $i=1,2,\ldots,n$.
$P(X_i)$	Probability that a dimensional random vector will take on observed values X_i.
ω_k	Member of the kth set of classes ω; $k=1, 2,\ldots, K$.
$P(\omega_k)$	Probability that an observation will be a member of class ω_k; prior probability of class ω_k.
$P(X\|\omega_k)$	Probability density value associated with an observation vector X as evaluated for class ω_k.
$p_k(X)$	Probability density value times prior probability for observation vector X evaluated for class ω_k.
μ_k	Parametric mean vector associated with the kth class.
Σ_k	Parametric p by p dispersion (variance-covariance) matrix associated with the kth class.
D_k	p by p dispersion matrix associated with a sample of observations belonging to the kth class; taken as an estimator of Σ_k.
$\sum_{i=1}^{n}$	Summation sign, add together all occurrences of i from 1 to n.
$\prod_{i=1}^{n}$	Product sign, multiply together all occurrences of i from 1 to n.
$\hat{\beta}$	Estimated vector of regression parameters.
m_k	Mean vector associated with a sample of observations belonging to the kth class; taken as an estimator of μ_k.

References

[1] Strahler, A.H. (1980). The use of prior probabilities in maximum likelihood classification of remotely sensed data; Remote Sensing of Environment, in press.

[2] Chow, C.K. (1957). An optimum character recognition system using decision functions. IRE Trans. Electron. Computers 6, pp. 247-254.

[3] Sebestyen, G. (1962). Decision-Making Processes in Pattern Recognition, MacMillan, New York.

[4] Nilsson, N.S. (1965). Learning Machines - Foundations of Trainable Pattern - Classifying Systems, McGraw-Hill Inc., New York.

[5] Schell, J.A. (1973), in Remote Sensing of Earth Resources, Volume I (F. Shahrokhi, Ed.), University of Tennessee Space Institute, Tullahoma, Tn., pp. 374-394.

[6] Reeves, R.G., Anson, A., and Landen, D. (1975), Manual of Remote Sensing, Amer. Soc. of Photogrammetry, Falls Church, VA, 2 vols., 2144 pp.

[7] Brooner, W.G., R.M. Haralick, and I. Dinstein (1971), Spectral parameters affecting automated image interpretation using Bayesian probability techniques: Proc. Seventh Intl. Symp. on Rem. Sens. of Env., pp. 1929-1948.

[8] Domenich, T.A. and McFadden, D. (1975). Urban travel demand: a behavioral analysis. Amsterdam, North-Holland.

[9] Cox, D.R. (1970). The analysis of binary data. London, Methuen.

[10] Mantel, N. and Brown, C. (1973). A logistic reanalysis of Ashford and Snowden's data on respiratory symptoms in British coal miners. Biometrics 29, pp. 649-65.

[11] Wrigley, N. (1975). Analyzing multiple alternative dependent variables. Geographical Analysis 7, pp. 187-95.

[12] Schmidt, P. and Strauss, R.P. (1975b). The prediction of occupation using multiple logit models. Int. Economic Review 16, pp. 471-86.

Nalepka, R.F., Horwitz, A.M., Hyde, P.D., Morgenstern, J.P. (1972), Classification of spatially unresolved objects. Manned Spacecraft Center, 4th Am. Earth Resources Program Rev., Vol. 2.

Woodcock, C.E., Smith, T.R., Strahler, A.H. (1979), a new model for estimating proportions of land cover within a pixel (abstr.), Machine Processing of Remotely Sensed Data Symposium.

Hartung, R.E., Lloyd, W.J. (1969), Influence of aspect on forests of the Clarksville soils in Dent County, Missouri. J. Forestry 67: 178-182.

Acknowledgements

This research has been supported by NASA grant NSG-2377. The authors would like to thank Joseph Scepan and Tara Torburn for the illustrations, and Debbie Heath and Kathy Bresslin for the typing.

INCORPORATING COLLATERAL DATA IN LANDSAT CLASSIFICATION AND MODELING PROCEDURES

A.H. Strahler, J.E. Estes, P.F. Maynard,
F.C. Mertz, D.A. Stow

Department of Geography
University of California
Santa Barbara, CA, 93106, U.S.A.

ABSTRACT

A number of existing statistical techniques can be used to merge spectral image data with collateral information, including regression, ANOVA, MANOVA, ANCOVA, MANCOVA, discriminant analysis, maximum likelihood classification with or without prior probabilities, contingency table analysis, and logit modeling. The choice of an appropriate technique depends upon the nature of input and output variables -- continuous, discrete, or categorical -- and the appropriate model -- parametric or nonparametric.

Logit modeling is a very versatile technique which is well adapted to remote sensing application. The logit, which is the natural logarithm of an odds ratio for two states of an output categorical variable, is predicted by a linear or curvilinear function of continuous or categorical input variables. Since the logit models probabilities or proportions, it can be used directly as a classifier or indirectly as an estimator of prior probabilities for conventional maximum likelihood classification with prior probabilities. The logit model is nonparametric, a feature which makes it highly desirable when used to merge disparate types of collateral data. The disadvantages of the logit model are that more calibration (training) data are required to fit the model, and that fitting requires an iterative nonlinear optimization procedure. A logit model was devised and fitted to forest species compositional data for northern California, predicting the proportion of timber volume in each of five species at each pixel based on registered terrain data quantifying elevation and slope aspect.

Another versatile tool is maximum likelihood classification with prior probabilities. By making prior probabilities conditional on a collateral data channel, the information contained within the channel can be conveyed to the maximum likelihood algorithm. An example is in land use, in which a previous classification and an externally devised transition probability matrix are used together with new image data to produce an updated classification consistent with the observed pattern of change. This technique has relevance for monitoring urban expansion and the impact of forest clearing in developing nations.

1. INTRODUCTION

Viewed in a broad context, the problem of combining image data and spatial collateral data into a predicted output map or image is actually a problem of combining continuous and categorical variables in a modeling framework capable of producing continuous or categorical outputs. At the University of California, Santa Barbara, NASA-supported research (grant NSG-2377) is currently underway to identify existing models and procedures for spatial modeling and to apply them to selected datasets to demonstrate their applicability in a remote sensing situation.

Collateral data, here defined as preexisting spatial information in the form of a map, pro-

cessed image, or set of observations at grid coordinate locations, can be combined with Landsat or other remotely sensed digital imagery to enhance classification accuracy or to construct models which predict spatial patterns of ground phenomena. Collateral information can be either continuous or categorical in nature. Examples are elevation, slope, or aspect channels obtained from a digital terrain model (continuous); and rock type, soil type, crop type, or land use (categorical). Image data, with which collateral information are to be combined, are typically continuous in nature, although values may be quantized into discrete gray tone levels for data processing applications. Desired outputs may also be either categorical or continuous. Any type of classification constitutes a discrete or categorical output, whereas such outputs as percent bare ground, timber volume, soil loss, or forage cover are continuous in character.

Figure 1 presents appropriate techniques for the merging of continuous, categorical and mixed (both continuous and categorical) datasets to produce either continuous or categorical outputs. Continuous output models, including regression, analysis of variance, and analysis of covariance, are all mathematically related and based on least squares algebra. Categorical output models are more diverse and include variance maximizing techniques such as discriminant analysis as well as the nonparametric methods of contingency table analysis and logit modeling. Maximum likelihood classification may be viewed as a special case of logit modeling in which the input variables are assumed to be normally distributed. Nonparametric and maximum likelihood techniques, because they produce probabilities of classification as an output, also have the advantage that they can be adjusted for prior probabilities. Because many present remote sensing applications call for categorical classification, these latter methods are probably most useful in combining continuous image data with categorical and continuous types of collateral data.

Demonstrations of a selected set of these techniques are planned and under current development; their current status is discussed in following sections. Several of these examples have important implications for remote sensing in developing nations. In one application, logit modeling is used to fit a model which describes the probability of occurrence of various forest species given elevation and slope aspect values obtained from a registered digital terrain model. Once obtained, these probabilities can be used as prior probabilities in a maximum likelihood classification of a Landsat image with registered terrain data for natural vegetation units. This technique could facilitate the accurate identification of forest types in complex tropical upland environments.

In another example, land use classification for change detection monitoring is improved by a contingency table-analytic technique which quantifies the probabilities of change for each land use type during the fixed time interval. This technique has important implications for many developing nations, especially in Central America, where urban expansion is impacting agricultural land, and forest clearing for agriculture is impacting large natural environments. Additional examples are being developed, focused on Landsat and collateral datasets obtained for an area of Ventura County, California.

2. STATISTICAL TECHNIQUES

Figure 1 identifies a set of statistical techniques which are relevant for combining collateral data in the context of Landsat modeling and classification. The techniques can be seen as a double level hierarchy, ranging from continuously measured independent variables in the first column to categorically measured independent variables in the last column. In the first row, the dependent variables are are continuous in nature, and in the second row the dependent variables are categorical. Continuous variables are measured on interval or ratio scales, whereas categorical variables are measured at nominal or ordinal scales. Categorical variables can be of three types:

(a) dichotomous (e.g., presence or absence, yes or no)
(b) unordered polychotomous (e.g., land uses; agriculture, urban, forest, etc.)
(c) ordered polychotomous (e.g., low runoff, medium, and high runoff)

The statistical techniques in the first row have been thoroughly documented and are commonly used in social science (Blalock, 1972; Graybill, 1961; Morrison, 1967; Winer, 1971). However, there has been relatively little work in the area of applying these approaches to remote sensing. The second row of the figure describes techniques which are generally less well known, but also

include maximum likelihood calssification as commonly carried out in remote sensing. The remaining portions of this section discuss the theory and remote sensing application of the techniques identified in Figure 1.

2.1 CONTINUOUS RESPONSE VARIABLE

The statistical techniques of the first row of Figure 1 share one thing in common -- they are all different forms of the basic linear regression model. Consequently, the theory and application of the different types of regression in row one will be very similar. Cell (a) will be examined in detail, and except where explicitly specified, the analysis can be extended to cells (b) and (c).

2.1.1 CELL (a) -- CONTINUOUS EXPLANATORY VARIABLES. The most commonly used method in this cell is regression, which predicts a continuously measured response variable from continuously measured explanatory variables. The model (in vector notation) can be written:

$$\underset{\sim}{Y} = \underset{\sim}{\beta}\underset{\sim}{X} + \underset{\sim}{\varepsilon}$$

where $\underset{\sim}{Y}$ is the vector of observed dependent variables, β is the vector of unknown parameters, $\underset{\sim}{X}$ is the vector of observed independent variables, and ε is the vector of errors. (Table I describes the symbols used in this paper.) Typically, the vector β is unknown, and must be estimated from a dataset for which both response and explanatory variables are observed. This estimation is done by the process of ordinary least squares regression (OLS). OLS regression finds the slope and the intercept of a line running through the data which minimizes the variance of $\Sigma(Y-\hat{Y})^2$, or the sum of the squared differences between the observed Y and the predicted $\hat{Y}$. The vector of predicted $\hat{Y}$ value is calculated by:

$$\underset{\sim}{\hat{Y}} = \underset{\sim}{\hat{\beta}}\underset{\sim}{X}$$

The regression is accomplished by defining the quantity Q equal to $\Sigma (Y - \hat{Y})$ taking the first partial derivatives of Q with respect to the values in the vector β and setting these partials equal to zero. By definition of a partial derivative, a minimum has been found.

The best statistic for measuring the strength of the regression is R^2. There are several ways of calculating R^2, but the following is conceptually the simplest. The residual sum of squares (RSS) or the total amount of variance in the model not explained by the regression is calculated by:

$$RSS = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

The RSS, when divided by the total corrected sum of squares in Y (written TSS), gives the proportion of unexplained variance to total variance. R^2, or the proportion of explained variance, is obtained by subtracting this ratio from one:

$$R^2 = 1 - \frac{RSS}{TSS}$$

One example of how regression could be used within the context of merging Landsat and collateral data is biomass modeling. The regression example used for cell (a) is restricted to continuously measured independent variables. For simplicity, only one spectral channel and one collateral data source are used. Adding other channels and other collateral data sources is a straightforward procedure. A basic linear model could be:

$$\hat{B}_i = \hat{\beta}_0 + \hat{\beta}_1 (MSS_i) + \hat{\beta}_2(rain_i)$$

where $\hat{B}_i$ is the predicted biomass for pixel i, MSS_i is observed multispectral scanner data (as single band, or multiband ratio or transform) for pixel i, $rain_i$ is observed rainfall on pixel i, and the vector of betas ($\hat{\beta}_0 \ldots \hat{\beta}_2$) are the estimated regression parameters.

Regression models are applied as a two stage process. First, the model must be calibrated (estimate the vector of $\hat{\beta}$'s) by regressing the observed independent variables on the observed

dependent variable. In this example, biomass data are required from a sufficient number of pixels to give a representative sample of the Landsat image to be modeled. The locations of data points are then "rubber-sheeted" to a geometrically corrected Landsat image, and the linked biomass and MSS data is directly accessed by a statistical software package, such as the Statistical Analysis System (SAS), to calculate the betas and R^2. Secondly, if R^2 is statistically significant (i.e., the calculated betas explain a significant portion of the total variance in Y), the model is extended as a predictor to other pixels where the dependent variable has not been observed. This, in effect, constitutes a model of biomass predicted by surface reflectance and measured rainfall.

2.1.2 CELL (b) -- MIXED EXPLANATORY VARIABLES. This cell includes conventional regression models which are similar to those of cell (a) but also include a mixture of continuous and categorical explanatory variables. Such models are common in social science research, the categorical explanatory variables often being termed "dummy" variables. It can be shown that the more familiar statistical test Analysis of Covariance (ANCOVA) is a straightforward extension of OLS regression with dummy variables (variables that assume values of 1 or 0 depending on the presence or absence of the qualitative variable being measured).

As an example, the model used in cell (a) will be extended to cell (b). In this cell we are able to include data measured categorically -- for example, soil type which can be observed in two states, referred to as class 1 and class 2. The model now becomes:

$$\hat{B}_i = \hat{\beta}_0 + \hat{\beta}_1(MSS_i) + \hat{\beta}_2(rain_i) + \hat{\beta}_3 D_{i1} + \hat{\beta}_4 D_{i2}$$

where the biomass for pixel i is predicted by the variables used in cell (a) in combination with the dummy variable terms $\hat{\beta}_4 D_{i1}$ and $\hat{\beta}_4 D_{i2}$. If the observed soil type for pixel i is class 1, then the categorical beta that will be used is $\hat{\beta}_3$, whereas if the soil type is class 2, then $\hat{\beta}_4$ will be used. This is accomplished by defining D_{i1} equal to 1 if the soil type in pixel i is class 1 and 0 otherwise and by defining D_{i2} equal to 1 if the soil type is class 2 and 0 otherwise. It is a relatively simple task to expand this model to include several (polychotomous) soil types or to include other categorical data.

When there is more than one interval scale dependent variable, the model is called MANCOVA, or Multivariate Analysis of Covariance. In this model, all of the independent variables are regressed against each of the dependent variables, with separate R^2s and F ratios calculated for each dependent variable. An example of this is:

$$\hat{B}_i, \hat{S}_i = \hat{\beta}_0 + \hat{\beta}_1(MSS_i) + \hat{\beta}_2(rain_i) + \hat{\beta}_3 D_{i1} + \hat{\beta}_4 D_{i2}$$

where $\hat{B}_i$ is predicted biomass for pixel i, $\hat{S}_i$ is predicted soil loss for pixel i, and the other variables remain as in the preceding example. MANCOVA can be seen as a device to test more than one ANCOVA model in the same statistical analysis.

2.1.3 CELL (c) -- CATEGORICAL EXPLANATORY VARIABLES. The extended regression model examined in cell (b) is also applicable here. When the explanatory variables are all measured on the ordinal or nominal scales, the model is usually called Analysis of Variance (ANOVA). All that is necessary to move from cell (b) to (c) is to exclude from the analysis all explanatory variables that are measured on the interval or ratio level. As an example of ANOVA, the biomass model would be written:

$$\hat{B}_i = \hat{\beta}_0 + \hat{\beta}_1 D_{i1} + \hat{\beta}_2 D_{i2} + \hat{\beta}_3 D_{i3} + \hat{\beta}_4 D_{i4}$$

Here, $\hat{\beta}_1$ and $\hat{\beta}_2$ refer to two different soil types and $\hat{\beta}_3$ and $\hat{\beta}_4$ refer to rainfall that has been categorized into two levels (high and low). It would be possible to categorize MSS data for use in such a model, but there is considerable information loss. Consequently, the ANOVA or MANOVA (Multivariate Analysis of Variance) model is not likely to be as useful as the ANCOVA or MANCOVA model.

2.2 CATEGORICAL RESPONSE VARIABLE

The first three cells have dealt with Landsat MSS data and continuous dependent collateral variables as predictors in various statistical models of physical phenomena. The last three

cells, in which the dependent variable is categorical, open up new ways of utilizing collateral data within the Landsat structure. With categorical data and with the appropriate statistical techniques (see Figure 1) it is possible to:

1) model physical and social phenomena that are best represented in discrete steps -- low, medium, high (soil erosion, fire hazard, biomass, housing quality, municipal services, etc.);
2) classify the dependent variable into nominal groupings (land use, vegetation community type, etc.);
3) create predicted probabilities that the dependent variable will assume particular categories and use these probabilities to classify an image directly or use them in conjunction with other data (usually MSS) in a Bayesian maximum likelihood classifier.

Unlike the first row of Figure 1, the second row includes five different statistical techniques. For purposes of modeling with Landsat under the constraints of the second row, Maximum Likelihood Classification (MLC) with Prior Probabilities is the most important method. "Maximum likelihood" is a statistical property of an estimator, and, used in its proper way, implies that an estimator has the highest probability of producing the data which were used for calibration. However, its use in remote sensing implies a particular decision rule (MLC) possessing this property, which is discussed below. Our research has shown that the inclusion of collateral data as prior probabilities to MLC offers a simple and effective way of combining collateral and remotely sensed data.

The key to the use of prior probabilities is the logit regression model, which takes collateral data (continuous or categorical) and generates predicted probabilities. These predicted probabilities can be used as a direct classifier (i.e., the pixel is classified into the category that has the highest probability), but the most likely usage of these predicted probabilities is as input to the MLC decision rule, in which they serve as weights. Since the logit regression model is probably the least familiar of the statistical techniques to remote sensing research, and since it is applicable in every cell in the bottom row, it will be explored with the most detail.

Discriminant Analysis is similar in its usage to maximum likelihood with prior probabilities but because of its computational complexity it has not been used often in the remote sensing context. Consequently, it will be only briefly discussed. Chi-Square Analysis, which is the traditional analysis used on contingency tables (all variables are categorical) is by definition not compatible with interval or ratio scale remotely sensed data. It is similar to ANOVA, except that the output is categorical.

2.2.1 CELL (d) -- CONTINUOUS EXPLANATORY VARIABLES. In the past ten years, maximum likelihood classification has found wide application in the field of remote sensing. Based on multivariate normal distribution theory, the MLC algorithm has been in use for applications in the social sciences since the late 1940's. Providing a probabilistic method for recognizing similarities between individual measurements and predefined standards, the algorithm found increasing use in the field of pattern recognition in the following decades (Chow, 1957; Sebestyen, 1962; Nilsson, 1965). In remote sensing, the development of multispectral scanning technology to produce layered multispectral digital images of land areas from aircraft or spacecraft provided the opportunity to use the maximum likelihood classifier in producing thematic classification maps of large areas for such purposes as land use/land cover determination and natural cultivated land inventory (Schall, 1972; Reeves et al., 1975).

Before presenting a practical example, it will be helpful to briefly review the mathematics of the maximum likelihood decision rule. In the multivariate remote sensing application, it is assumed that each observation X (pixel) consists of a set of measurements on p variables (channels). Through some external procedure a set of observations which correspond to a class is identified -- that is, a set of similar objects characterized by a vector of means on measurement variables and a variance-covariance matrix describing the interrelationships among the measurement variables which are characteristic of the class. Although the parametric mean vector and dispersion matrix for the class remain unknown, they are estimated by the sample means and dispersion matrix associated with the object sample.

Multivariate normal statistical theory describes the probability that an observation X will

occur, given that it belongs to a class k, as the following function:

$$\phi_k(X_j) = (2\pi)^{-\frac{1}{2}p} |\Sigma_k|^{-\frac{1}{2}} e^{(X - \mu_k)' \Sigma_k^{-1} (X - \mu_k)}$$

As applied in a maximum likelihood decision rule, the previous expression allows the calculation of the probability that an observation is a member of each of k classes. The observation (pixel) is then assigned to the class for which the probability density value is greatest. Since the log of the probability is a monotonic increasing function of the probability, the decision can be made by comparing values for each class as calculated from the right hand side of equation.

A simplified remote sensing classification rule using maximum likelihood (Tatsuoka, 1971; Strahler, 1980) with k possible categorical classes and p channels of MSS input datasets is to choose the k (class) which minimizes

$$F_{1,k}(X_j) = \ln|D_k| + (X_j - m_k)' D_k^{-1} (X_j - m_k).$$

This expression is derived from the preceding one by taking the natural logarithm and deleting terms which are constant for all classes.

Interval or ratio level collateral data can be incorporated as extra "logical" channels within this model. One successful forestry application was achieved by the creation of a texture channel which was synthesized from Landsat Band-5 by taking the standard deviation of density values within a 3 by 3 moving window, scaling this value, associating it with the center pixel of the 3 by 3 window, and returning it in image format (as a fifth channel). Values in the texture channel describe the variation in image tone within the immediate area of each pixel. High values are characteristic of edges and boundaries, whereas lower values describe more uniform areas. This technique was shown to significantly increase classification accuracies for a number of species-specific forest cover types in northern California (Strahler, 1978, 1979). Strahler (1978) demonstrated how to input collateral information in the form of elevation data and slope aspect (in combination with a texture channel) as separate "logical" channels, increasing the classification accuracy by 27 percent.

2.2.1.1 Logit Regression. In extending the conventional regression models adopted in cells (a), (b), and (c) to the problems of cell (d), two difficulties are encountered. (For details see Wrigley, 1976, p. 8-9; 1977b; p. 12-13). First, a conventional regression model with a categorical response variable will violate the constant error variance or homoscedasticity assumption. While this problem does not result in biased or inconsistent parameter estimates, it does result in a loss of efficiency and gives rise to serious problems if conventional inferential tests are used. Secondly, a conventional regression model with a categorical response variable may generate predictions which are seriously deficient. It can be shown that the predicted values of the response variable in such a model are best interpreted as predicted probabilities. The problem is that although probabilities are constrained to lie within the range of 0 to 1, the predictions generated from such a model are unbounded and may take values from minus infinity to plus infinity. Thus, the predictions may lie outside the meaningful range of probability and may be inconsistent with the probability interpretation that was just presented. The simplest yet most statistically sound solution to the probability problem (within a regression framework) is the logit transformation, in which the probability P_i is modeled as

$$P_i = \frac{e^{\hat{\beta}X}}{1 + e^{\hat{\beta}X}}$$

and the probability of "not P_i" is:

$$1 - P_i = \frac{1}{1 + e^{\hat{\beta}X}}$$

where $\hat{\beta}X$ is the vector product of betas multiplied by row vector of X's (observed explanatory variables). Although these two equations are nonlinear models, it is a simple matter to rewrite them as

$$\frac{p_i}{1 - p_i} = e^{\hat{\beta}X}$$

The logit transformation is achieved by taking the natural logarithm of the preceding formula, which yields

$$\ln \frac{p_i}{1 - p_i} = \hat{\beta}X$$

This transformation has the property of increasing from minus infinity to plus infinity as p_i increases from 0 to 1. Once efficient estimators are calculated for the betas, simple algebra will extract the value of p_i. The method can be generalized to k classes, in which there are $k - 1$ logits of the form

$$\ln(\frac{p_2}{p_1}),\ \ln(\frac{p_3}{p_1}),\ldots,\ln(\frac{p_k}{p_1})$$

Each logit must be modeled separately, producing $k - 1$ sets of betas. As in the binary case described above, algebra will extract the values of the probabilities from the $k - 1$ logits predicted for an observation along with the constraint that all probabilities must sum to one.

Unlike the conventional regression models of the previous cells, all of which can be efficiently estimated by the ordinary least squares (OLS) method, the logistic and linear logit regression models appropriate for the problems of cell (d) require either a weighted least squares (WLS) estimation procedure or a maximum likelihood procedure. The choice between the two methods depends upon whether the calibration data include repeated observations for each combination of values of the explanatory variables (in the case of WLS) or not (in the case of maximum likelihood). Since such replications are unlikely in calibration of the logit model for remotely sensed data, maximum likelihood estimation is preferred. (Note that here the term "maximum likelihood" refers to a parameter estimation method different from multivariate normal MLC.) The maximum likelihood solution to the calibration of the logit model has many attractive features. It can be shown that provided the sample data are not multicollinear, a unique maximum likelihood estimator can be obtained even in relatively small samples. Also, the mathematical properties of the likelihood function allow for efficient computer programs to produce the parameter estimates, and these estimates are consistent and are the best possible estimates in very large samples. The disadvantages of the procedure are that it involves numerical optimization and therefore more costly computation, and that it is a less familiar statistical technique. For a description of the procedure used to calculate such maximum likelihood estimates, see Cox (1970, p. 87), Mantel and Brown (1973, p. 654-5), Wrigley (1975, p. 191-3), Domencich and McFadden (1975, p. 110-12) and Schmidt and Strauss (1975a, p. 484-5).

There are two ways the logit model can be used within the context of remote sensing. The first is to act directly as a nonparametric classifier (i.e., to classify a pixel into the class having the highest predicted probability). The second is to use collateral data in a logit model to predict prior probabilities and input them to a MLC decision rule which accepts prior probabilities. This approach effectively combines a nonparametric logit model for collateral data with a parametric MLC model; it will be discussed in a following section. The following example of the logit model involves the direct classification of land use by MSS and terrain data. The model is

$$\ln \frac{p_{i1}}{1 - p_{i1}} = \hat{\beta}_0 + \hat{\beta}_1(band5_i) + \hat{\beta}_2(slope_i)$$

where p_{i1} is the probability that pixel i is class 1 and the following terms indicate linear combination of spectral and collateral data. If desired, the model can easily be expanded to all four bands and all the continuous collateral variables relevant to the classification.

2.2.1.2 Discriminant Analysis. Discriminant analysis is a multivariate technique used to produce sets of uncorrelated functions which separate observations most efficiently into predesignated groups. A discriminant model, or classification criteron, is developed, the values of

which define groups for the observations. The individual observation is classified into one of the previously defined groups by a measure of generalized squared distance.

This technique requires some difficult computation. Deriving the classification criteria requires extracting eigenvectors from the nonsymmetric $W^{-1}B$ matrix, where B and W are respectively, the between and within groups sums of squares and crossproducts matrices. The mathematics of this process are beyond the scope of the paper (please refer to Tatsuoka, 1971; Cooley and Lohnes, 1971). The technique is helpful in the social sciences for identifying variables which do the best job of separating classes, but is not usually used to process new data for classification. Further, the technique assumes that all classes possess an identical dispersion matrix, an assumption unlikely to characterize remotely sensed data.

2.2.2 CELL (c) -- MIXED EXPLANATORY VARIABLES. Conceptually no new problems are encountered in moving from cell (d) to cell (c); categorical explanatory variables are included through dummy variables. The logit model has high potential for application in remote sensing. Since it is capable of incorporating both bontinuous and discrete input data and generating probabilities either directly for classification or as input as a collateral data set of probabilities to MLC (Strahler, 1979). In other words, interval level measured data such as rainfall, elevation, slope, etc. (no matter what its variance-covariance) can be combined with discrete data such as soil type, previously classified land use, census tracts, etc., in a nonparametric, logit framework and the result will be a discrete output such as a land use map, a soil erodibility map (divided into discrete levels) or other special use maps.

The logit model, as in the previous cell, can be easily extended to cell (c). Again, land use could be predicted by

$$\ln \frac{P_{i1}}{1 - P_{i1}} = \hat{\beta}_0 + \hat{\beta}_1(band5_i) + \hat{\beta}_2(slope_i) + \hat{\beta}_3 D_{i1} + \hat{\beta}_4 D_{i2}$$

where P_{i1} is the probability of land use '1' for pixel i over the probability of all the land uses that are not '1', $band5_i$ is the MSS value for pixel i, $slope_i$ is slope of the pixel, and D_{i1} is a dummy variable with a value of 0 if the previous land use on pixel i was class 1 and 0 otherwise and D_{i2} has the value 1 if the land use on pixel i was class 2 and 0 otherwise.

Preceeding paragraphs have referred to the use of prior probabilities in modifying the outcome of a MLC. Since the prior probabilities can be modeled to reflect both continuous and categorical input data, and the input of MLC is a categorical calssification, this tehcnique is appropriate to discussion of cell (c). Prior probabilities are incorporated into the classification through the manipulation of the Law of Conditional Probability. The acutal derivation of the prior probability is beyond the scope of this paper (see Strahler, 1980). The modified decision rule is to choose k which minimizes

$$F_{2,k}(X_j) = \ln|D_k| + (X_j - m_k)' D_k^{-1} (X_j - m_k) - 2\ln P(\theta_k)$$

where the only difference between this formula and the one presented in cell (d) is the probability term, $-2\ln P(\theta_k)$. This form of the decision rule is usually attributed to Tatsuoka and Tiedeman (1954; Tatsuoka, 1971).

It is important to understand how this decision rule behaves with different prior probabilities. If the prior probability $P(\theta_k)$ is very small, then its natural logarithm will be a large negative number; when multiplied by -2, it will become a large positive number and thus $F_{2,k}$ for such a class will never be minimal. Therefore, setting a very small prior probability will effectively remove a class from the output classification. Note that this effect will occur even if the observation vector X_j is coincident with class mean vector m_k. In such a case, the quadratic product distance function $(X_j - m_k)' D_k^{-1} (X_j - m_k)$ goes to zero, but the prior probability term $-2\ln P(\theta_k)$ can still be large. Thus it is entirely possible that the observation will be clsssified into a different class, one for which the distance function is quite large.

As the prior probability $P(\theta_k)$ becomes large and approaches 1, its logarithm will go to zero and $F_{2,k}$ will approach $F_{1,k}$ for that class. Since this probability and all others must sum to one, however, the prior probabilities of the remaining classes will be small numbers and their

values of $F_{2,k}$ will be greatly augmented. The effect will be to force classification into the class with high probability. Therefore, the more extreme are the values of the prior probabilities, the less important are the actual observations values X_j.

For a numerical example of how prior probabilities can affect the decision of the maximum likelihood classifier, please refer to Strahler, 1980. There are so many potential applications of prior probabilities and the maximum likelihood decision rule that it would be counterproductive to list them all. In general, *all* data that is relevant to a classification model can now be incorporated and this process has been shown to significantly increase classification accuracies (Strahler, 1978; Strahler, 1980).

The versatility of the prior probability techniques comes about when the priors are allowed to vary on a pixel-by-pixel basis. The priors for a pixel may be determined by a logit or other model, or by using a set of class-conditional prior probabilities estimated by sampling. Because the priors are computed separately, it is possible to mix any sort of model estimating prior probabilities with a multivariate normal MLC algorithm which is known to be well suited to most spectral data. Thus, the technique allows easy, flexible merging of collateral data, used to predict the priors, with continuous image data. These points are discussed in more length in Strahler (1978).

2.2.3 CELL (f) -- CATEGORICAL EXPLANATORY VARIABLES. Data that falls into this level has been traditionally analyzed by contingency table analysis with Chi-Square methods. But statistically speaking, it is a simple matter to extend the logit model of cell (e) to cell (f) through the use of dummy variables. For data that only come in nominal or ordinal levels, the logit model offers new and important insights into the data (Wrigley, 1979; Theil, 1970; Grizzle, Starmer and Koch, 1969; Koch et al., 1971, 1972, 1976a, 1977; Landis and Koch, 1977; Lahman and Koch, 1974a, 1974b). As in earlier discussion, the logit model can serve directly as a nonparametric classifier, using only categorical variables input as dummy variables. In this form, the logit model is equivalent to a log linear model of a contingency table; such models are discussed fully in such texts as Bishop et al. (1975).

The categorical logit model is formulated in the example below:

$$\ln \frac{P_{1i}}{1 - P_{1i}} = \hat{\beta}_0 + \beta_1 D_{1i} + \hat{\beta}_2 D_{2i}$$

where there are two output categories -- 1 and not 1 -- which are modeled by categories of soil type as described in cell (e). The classifier simply assigns the output pixel to the class with the higher probability.

3. APPLICATIONS

Two statistical modeling techniques, logit modeling and maximum likelihood classification with prior probabilities, were selected for further investigation in the context of a real application. A linear logit model was devised and fitted to forest species compositional data for northern California, predicting the proportion of timber volume for each of five coniferous species at each pixel based on registered terrain data quantifying elevation and slope aspect. Maximum likelihood classification with prior probabilities was tested in Ventura County, California in a land use application. A previous Landsat classification and an externally devised transition probability matrix were used together with new Landsat image data to produce an updated classification consistent with the observed pattern of change.

Throughout this research we have utilized the Video Image Communication and Retrieval (VICAR) system and the Image Based Information System (IBIS) resident at UCSB. VICAR/IBIS, developed at the Jet Propulsion Laboratory (JPL) at Pasadena, California, USA, is a job control language which permits the sequential linking and execution of a vast array of Fortran and Assembler routines in a batch environment. In addition to extensive usage of existing VICAR/IBIS routines, new VICAR and non-VICAR software were developed and/or modified as required for the purposes of this research.

3.1 LOGIT MODELING OF SPECIES PROPORTIONS

Logit modeling of species proportions used data derived from the Klamath National Forest, located in northern California, USA, (Figure 2). Ranging in relief from 500 to 8,000 feet, the Forest includes 2,600 square miles of rugged terrain in the Siskiyou, Scott Bar, and Salmon Mountains. Little of the area is developed beyond management for timber yield, livestock production, and recreation. A wide variety of distinctive vegetative types is present in the area. Forest vegetation includes such coniferous species as noble, red, white, and douglas firs, ponderosa pine, and incense cedar, as well as several oaks, and typical species of chaparral. Thus, the topographic and vegetational characteristics of the area are well differentiated. Within the Klamath National Forest, a study area including most of the Goosenest Range was selected for logit modeling of species composition from terrain features. This area was chosen because calibration data and Landsat images were readily available for it.

The logit model devised for this forestry application requires preparation of digital terrain data. These data, obtained from the National Cartographic Information Center, in Reston, Virginia, USA, are derived from processing of 1:250,000 contour maps, and include elevations at every point on a grid of approximately 65 m spacing. Although the data are comparable in scale to a Landsat image, the elevation values are quite generalized because they are produced from small scale contour maps by interpolation (Figure 4).

Slope angle and slope aspect channels can be produced using the elevation data of the registered terrain image. Although a number of slope and aspect generating algorithms are known, the simplest is the fitting of a least squares plane through each pixel and its four nearest neighbors and the calculation of the downslope angle and direction of the plane. Slope angle is obtained relative to the numeric range of the elevation channel and image grid spacing, and azimuth is determined with respect to the rectangular image grid. These channels are best generated directly during the preprocessing of the original terrain data. At that time, slope angles and aspects can be calculated from half-word absolute elevations arrayed in a north-south east-west grid.

The slope aspect image (Figure 5) consisted initially of gray tone densities between 0 (black) and 255 (white) which indicated the azimuth of slope orientation, ranging clockwise from 0° to 359°. These values were then transformed according to the function below:

$$newden = 3.0 + 126.0*(1.0 + \cos(.024933275*(olden - 26.1)))$$

where *olden* symbolizes the old (azimuth-keyed) gray tone pixel density value, *newden* symbolizes its transformed value, and the argument of the cosine function is expressed in radians. This function transforms density values according to an orientation proposed by Hartung and Lloyd (1969). Since northeast slopes present the most favorable growing environment, and southwest slopes the least favorable, with northwest and southeast slopes of neutral character, the density tone azimuths were rescaled by a cosine function with 3 representing due northeast and 255 representing due southwest. Neutral slopes, oriented northwest or southwest, thus received density tones near 128. Flat pixels were coded with zeroes. The function also corrects automatically for the 12° skew of the Landsat image.

The logit model fitted is shown below:

$$\ln(\frac{P_k}{P_{-k}}) = \hat{\beta}_{1,k} + \hat{\beta}_{2,k}E + \hat{\beta}_{3,k}A + \hat{\beta}_{4,k}S, \; k = 1,5$$

where P_k is the probability that the board-foot of timber volume will be drawn from one of five species k, P_{-k} is the probability that the board-foot will not be drawn from species k, E is elevation (compressed to 0-255 range), A is aspect transformed as described above, S is slope angle, and $\hat{\beta}_{1,k},\ldots,\hat{\beta}_{4,k}$ are the estimated regression constants. Note that five equations, one for each species, actually comprise the model. The model was calibrated using 73 measurements of timber volume prepared by the U. S. Forest Service and located within two subregions of the Goosenest range. These samples are probably not representative of the entire area modeled, but serve for the demonstration purposes of this research. Each sample was located on 1:15,840 scale color air photos and tranferred to Band 5 of the Landsat images to obtain the line and sample coordinates of the sample point. The coordinates were then used to extract elevation and aspect values for the sample from the registered elevation and aspect images. The coefficients for the model were fitted by a nonlinear optimization algorithm employing the Newton-Raphson method to select

coefficients with maximum likelihood.

Given the constants produced by the procedure discussed above, the probability images were created using the new VICAR program "PROBMAPS." PROBMAPS, written specifically for this application, calculates the probability of species k for each pixel using the following expression:

$$P_k = \frac{Q_k}{\sum_{j=1}^{5} Q_j} \quad \text{where } Q_k = \exp\left(\hat{\beta}_{1,k} + \hat{\beta}_{2,k}E + \hat{\beta}_{3,k}A + \hat{\beta}_{4,k}S\right)$$

PROBMAPS then scales each probability so that the range 0-255 represents 0. to 1. PROBMAPS output images for this example are shown in Figures 6-10. Brightness values in Figures 6-9 represent probabilities of occurrence for douglas fir, ponderosa pine, white fir, and red fir, respectively, with probabilities scaled to range from black (0.) to white (1.0). Figure 10, incense cedar, has been contrast stretched for display purposes, and presents a probability range of 0. to .3 from black to white. The probability images represent maximum likelihood estimates of species proportions; they appear reasonable in light of the known ecological preferences of the species, but their accuracies remain to be determined.

The probability images produced by PROBMAPS can be thought of as predictions of the proportion of the timber volume expected for each species at each pixel. This view implies a continuous mixture of species, constatly varying in response to elevation and slope aspect. An alternative view is that forest stands are monospecific, and that each pixel is dominated by a particular species or forest cover type. In such a case, the modeled values are probabilities that the pixel will be dominated by a particular species or stand type, and it is therefore appropriate to produce a single output image indicating the type with highest probability for each pixel. In this way, the logit model can serve as a nonparametric classifier. The probabilities can also be viewed as prior probabilities, and input to MLC of an image using spectral data. This procedure amounts to mixing a nonparametric model for collateral data (terrain channels) with a parametric model for spectral data (Landsat channels).

3.2 LAND USE CLASSIFICATION USING TRANSITION PROBABILITIES

An additional objective of our research was to apply the method of maximum likelihood classification with prior probabilities to a land use/land cover classification (see sections 2.2.1 and 2.2.2). In this example, land use/land cover maps for two 7½-minute quadrangles in Ventura County, California, U.S.A., were obtained from photointerpretation of high-level U-2 aircraft imagery for the years 1973 and 1976. These maps, with inherent accuracies considerably higher than those of Landsat classifications, were used as "ground truth" to construct a matrix of transition probabilities, showing the probability of change of classification for each land use/land cover type to each other type in the three year interval. With a 1976 Landsat image as input data, we planned to carry out MLC of each pixel using the 1973 U-2 derived cover class as a collateral data channel indexing the transition probabilities appropriate to the 1973 cover class. The resulting classification was to be compared with the 1976 U-2 derived map to evaluate the accuracy of the technique. At present, we have used image overlay techniques to create the transition probability matrix, but the classifications using the transition probability matrix have not been carried out.

3.2.1. PROCEDURE. Using the Jet Propulsion Lab's (JPL) Image Based Information System (IBIS) and a coordinate digitizer, it is possible to merge image data in digital form with other types of geographic data. The IBIS is essentially a fine-mesh grid information system which is compatible with the handling and storing of digital image data. By allowing a user to overlay thematic map and digital Landsat data (or pertinent Landsat-derived data) with IBIS, it is possible to derive the values that comprise the transition probability matrix, as well as determine the accuracy of the thematic Landsat classification data.

Prior to IBIS processing, the two "ground truth" land use/land cover maps for the two-quadrangle study area were prepared through photointerpretation of NASA high-altitude color infrared imagery with additional ground checks. A coordinate digitizer board was used to convert the maps to a series of digital coordinates. Polygons of thematic land cover categories were captured by digitizing overlapping line segments that comprise such polygons. Polygon centroids were also

digitized and assigned an appropriate land use/land cover label for later use in converting the polygonal data to raster (image base) form.

The digitized line segment data were processed using IBIS as follows. A modified version of the IBIS program POLYGEN converted the coordinate digitizer segment data into polygons in the form of an IBIS graphics file. Following this reformatting, the program POLYREG rigidly rotated the polygons to conform with the Landsat data for the test site. POLYSCRB was used to convert the polygon data into raster form (fine-mesh grid). The result was an image of rasterized polygon borders representing the edges to thematic land cover units.

The next phase in the IBIS processing involved the assignment of labels to the rasterized polygon areas. An intermediate categorized image was automatically produced by the program PAINT, which assigns an arbitrary but unique brightness number to all the pixels within each rasterized polygon. This output image was combined with the polygon centroids that had also been converted to an IBIS graphics file format (with V2POLY) and rigidly transformed to overlay with the polygon borders (POLYREG). The program CTRWATCH was used to establish the correspondence between centroid labels and polygon brightness numbers, and the program STRETCH was used to reassign identical brightness values to polygons with similar labels, yielding a raster format image corresponding directly with the original land use/land cover map.

With both land use/land cover images in digital image format, the next step is to overlay them using POLYOVLY to analyze the change occurring between the two dates in order to estimate the transition probability matrix. The output of POLYOVLY is a table counting the number of pixels in each combination of classes across the two images. If $S_{i,j}$ denotes the count of pixels in class i of the first image, and class j of the second, then the transition probability $T_{i,j}$ for class i to j is $T_{i,j} = \frac{S_{i,j}}{\sum_{i=1} S_{i,j}}$. The digitized map of 1976 land use/land cover was also used to assess the accuracy of MLC of Landsat data, again using the POLYOVLY program. Accuracies discussed below were obtained in this fashion.

With a digitized 1973 land use/land cover classification map produced from air photo interpretation now in hand as collateral data channel registered to the 1977 Landsat image, and with a transition probability matrix to provide sets of prior probabilities contingent as a collateral data channel, it should be possible to carry out maximum likelihood classification using the transition probabilities as prior probabilities indexed by the 1973 classification. Future work includes performing the classification, and overlaying it on the 1976 digitized map for accuracy analysis. Initial indications are that accuracies will increase with the use of the 1973 digitized map data, demonstrating the successful use of Landsat to update existing manually produced land use/land cover maps.

4. CONCLUSION

Of the large number of statistical techniques which can be used to develop models combining remotely sensed images with collateral data in a common predictive framework, two techniques are of special interest for remote sensing: logit modeling and maximum likelihood classification with prior probabilities. These methods allow the construction of nonparametric classification models utilizing both image and collateral data channels as well as the mixing of parametric and nonparametric classification models for image and collateral data respectively. Both techniques have been successfully demonstrated in applications using Landsat imagery; each has the potential to greatly increase classification accuracy through the use of collateral data, and each should find wide application in future research and development in remote sensing.

5. REFERENCES

Bishop, Y.M., Fienberg, S.E., and Holland, P.W. (1975), Discrete Multivariate Analysis: Theory and Practice, MIT Press, Cambridge, Massachusetts.

Blalock, H.M. (1972), Social Statistics, McGraw-Hill Inc., New York.

Chow, C.K. (1957), An optimum character recognition system using decision functions, IREE Trans. Election. Computers 6: pp. 247-254.

Cox, D. R. The analysis of binary data. 1970. London. Methuen.

Domencich, T.A. and McFadden, D. Urban travel demand: a behavioral analysis, 1975, Amsterdam, North-Holland.

Graybill, F.A. (1961), An Introduction Linear Statistical Models, Vol.1, McGraw-Hill Book Co., New York.

Grizzle, J.E., Starmer, C.F. and Koch, C.G. Analysis of categorical data by linear models. Biometrics 25, 1969, pp. 489-504.

Hartung, R.F., Lloyd, W.J. (1969), Influence of Aspect on Forests of the Clarksville Soils in Dent County, Missouri, Journal of Forestry, 67, pp. 178-182.

Koch, G.G., Imrey, P.B. and Reinfurt, D.W. Linear model analysis of categorical data with incomplete response vectors. Biometrics 28, 1972, pp. 663-92.

Koch, G.G., Freeman, J. L. and Lehnen, R.G. A general methodology for the analysis of ranked policy preference data. International Statistical Review 44, 1976a, pp. 1-28.

Koch, G.G., Landis, J.R., Freeman, J.L., Freeman, D.H. and Lehnen, R.G. A general methodology for the analysis of experiments with repeated measurement of categorical data. Biometrics 33, 1977, pp. 133-58.

Landis, J.R. and Koch, G.G. The measurement of observer agreement for categorical data. Biometrics 33, 1977, pp. 159-74.

Lehnen, R.G. and Koch, G.G. A general linear approach to the analysis of nonmetric data: applications for political science. American Journal of Political Science 18, 1974a, pp. 283-313.

Lehnen, R.G. and Koch, G.G. The analysis of categorical data from repeated measurement research designs. Political Methodology 1, 1974b, pp. 103-23.

Mantel, N. and Brown, C. A logistic re-analysis of Ashford and Snowden's data on respiratory symptoms in British coal miners. Biometrics 29, 1973, pp. 649-65.

Morrison,, (1967), Multivariate Statistical Methods, McGraw-Hill Book Co., New York.

Nilsson, N.S. (1965), Learning Machines - Foundations of Trainable Pattern - Classifying Systems, McGraw-Hill Book Co., New York.

Reeves, R.G., Anson, A., and Landen, D. (1975), Manual of Remote Sensing, Amer. Soc. of Photogrammetry, Falls Church, VA, 2 vols., 2144 pp.

Schell, J.A. (1973), in Remote Sensing of Earth Resources, Volume I (F. Shahrokhi, Ed.), University of Tennessee Space Institute, Tullahoma, TN, pp. 374-394.

Schmidt, P. and Strauss, R.P. The prediction of occupation using multiple logit models. International Economic Review 16, 1975b, pp. 471-86.

Sebestyen, G. (1962), Decision-Making Processess in Pattern Recognition, MacMillan, New York.

Strahler, Alan H., T.L. Logan, and N.A. Bryant (1978). Improving forest cover classification accuracy from Landsat by incorporating topographic information: Proceedings of the Twelfth International Symposium on Remote Sensing of the Environment, pp. 927-942.

Strahler, Alan H., and T.L. Logan, and C.E. Woodcock (1979). Forest classification and inventory system using Landsat, digital terrain, and ground sample data: Proceedings of the Thirteenth International Symposium on Remote Sensing of the Environment, pp. 1541-1557.

Strahler, Alan H. (1980). The use of prior probabilities in maximum likelihood classification of remotely sensed data: submitted for publication.

Tatsuoka, M.M., Multivariate analysis: Techniques for Education and Psychological Research. 1971, John Wiley & Sons, New York.

Tatsuoka, M.M. and Tiedeman, D.V. Discriminiant Analysis Review of Education Research, 25, 1954, pp. 402-420.

Winer, B.J. (1971), Statistical Principles in Experimental Design, 2nd ed., McGraw-Hill Book Co., New York.

Wrigley, N. (1975), Analyzing multiple alternative dependent variables. Geographical Analysis 7, pp. 187-95.

Wrigley, N. (1976), An introduction to the use of logit models in geography. Concepts and techniques in modern geography, 10. Norwich: Geo Abstracts Ltd.

Wrigley, N., (1976b), Probability surface mapping: a new approach to trent surface mapping. Transactions of the Institute of British Geographers, New Series 2, pp. 129-40.

Wrigley, N. (1979), Development in the statistical analysis of categorical data -- a review. Progress in Human Geography, 3, pp. 315-355.

Table I. Notation

TERM	DEFINITION
$\hat{\underset{\sim}{Y}}$	Vector of estimated dependent variables; where $\sim$ signifies a vector and $\hat{}$ signifies an estimator.
$\hat{\underset{\sim}{\beta}},\underset{\sim}{\varepsilon}$	Regression notation; vector of estimated betas and vector of observed error terms.
p	Number of measurement variables used to characterize each object or observation.
X	A p-dimensional random vector.
X_i	Vector of measurements on p variables associated with the ith object or observation: $i=1,2,\ldots,N$.
θ_k	Member of the kth set of classes θ; $k=1,2,\ldots K$.
$P(\theta_k)$	Probability that an observation will be a member of class θ_k; prior probability for class θ_k.
$\phi_k(X_i)$	Probability density value associated with observation vector X as evaluated for class k.
μ_k	Parametric mean vector associated with the kth class.
m_k	Mean vector associated with a sample of observations belonging to the kth class; taken as an estimator of μ_k.
Σ_k	Parametric p by p dispersion (variance-covariance) matrix associated with the kth class.
D_k	p by p dispersion matrix associated with a sample of observations belonging to the kth class; taken as an estimator of Σ_k.

INPUT CHANNELS [Independent Variable(s)]

OUTPUT CHANNEL
[Dependent Variable]

	Continuous	Mixed	Categorical
Continuous	Regression Models - linear - curvilinear	Analysis of Covariance Multivariate Analysis of Covariance	Analysis of Variance Multivariate Analysis of Variance
Categorical	Maximum Likelihood Classification Logit Modeling Discriminant Analysis	Maximum Likelihood Classification with Prior Probabilities Logit Modeling	Contingency Table Analysis Logit Modeling

Figure 1. Techniques for Combining Continuous and Categorical Data (modified from Wrigley, 1979).

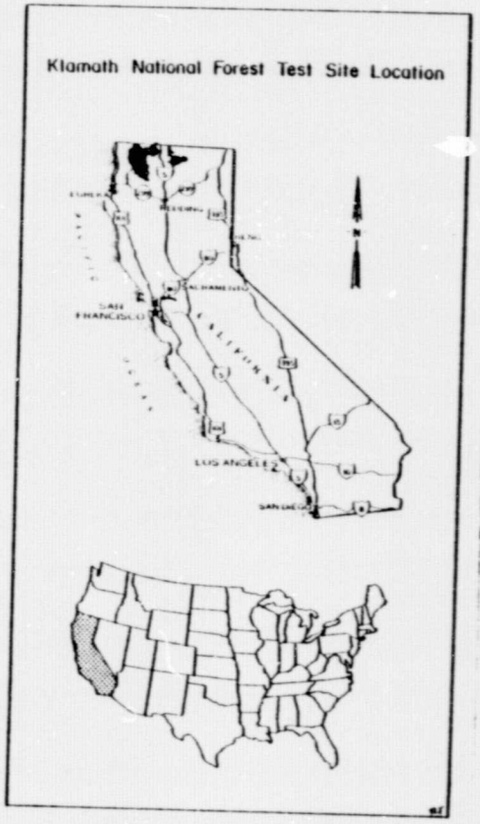

Figure 2. Klamath National Forest location map.

Figure 3. Landsat TM band 5, Goosenest test area, Klamath National Forest, California, U.S.A. Scale: pixel = 80 m x 80 m. Note: North is 9° counterclockwise from up.

Figure 4. Registered elevation image (derived from N.C.I.C. 1:250,000 digital terrain data).

Figure 5. Registered slope orientation image (cosine transformation).

Figure 6. Registered douglas fir probability image. (Area shown is restricted to National Forest lands).

Figure 7. Registered ponderosa pine probability image.

Figure 8. Registered white fir probability image.

Figure 9. Registered red fir probability image.

Figure 10. Registered incense cedar probability image (stretched for contrast enhancement).

The Use of Prior Probabilities in Maximum Likelihood Classification of Remotely Sensed Data

ALAN H. STRAHLER

University of California, Santa Barbara, California

The expected distribution of classes in a final classification map can be used to improve classification accuracies. Prior information is incorporated through the use of prior probabilities — that is, probabilities of occurrence of classes which are based on separate, independent knowledge concerning the area to be classified. The use of prior probabilities in a classification system is sufficiently versatile to allow (1) prior weighting of output classes based on their anticipated sizes; (2) the merging of continuously varying measurements (multispectral signatures) with discrete collateral information datasets (e.g., rock type, soil type); and (3) the construction of time-sequential classification systems in which an earlier classification modifies the outcome of a later one.

The prior probabilities are incorporated by modifying the maximum likelihood decision rule employed in a Bayesian-type classifier to calculate *àposteriori* probabilities of class membership which are based not only on the resemblance of a pixel to the class signature, but also on the weight of the class which is estimated for the final output classification. In the merging of discrete collateral information with continuous spectral values into a single classification, a set of prior probabilities (weights) is estimated for each value which the discrete collateral variable may assume (e.g., each rock type or soil type). When maximum likelihood calculations are performed, the prior probabilities appropriate to the particular pixel are used in classification. For time-sequential classification, the prior classification of a pixel indexes a set of appropriate conditional probabilities reflecting either the confidence of the investigator in the prior classification or the extent to which the prior class identified is likely to change during the time period of interest.

Introduction

In the past ten years, maximum likelihood classification has found wide application in the field of remote sensing. Based on multivariate normal distribution theory, the maximum likelihood classification algorithm has been in use for applications in the social sciences since the late 1940's. Providing a probabilistic method for recognizing similarities between individual measurements and predefined standards, the algorithm found increasing use in the field of pattern recognition in the following decades (Chow, 1957; Sebestyen, 1962; Nilsson, 1965). In remote sensing, the development of multispectral scanning technology to produce layered multispectral digital images of land areas from aircraft or spacecraft provided the opportunity to use the maximum likelihood criterion in producing thematic classification maps of large areas for such purposes as land use/land cover determination and natural cultivated land inventory (Schell, 1972; Reeves, et al., 1975).

In the last decade, research on the general use of classification algorithms in remote sensing has centered in two areas: (1) computational improvements in evaluating maximum likelihood and discriminant function decision rules; and (2) the use of various unsupervised clustering algorithms to extract repeated or commonly occurring measurement vectors which are characteristic of a particular multispectral scene. Computational improvements have included such developments as look-up table schemes (Schlien and Smith, 1975) to reduce repeated calculation, and hybrid classifiers (Addington, 1975) which use parallelepiped algorithms (Goodenough and Schlien, 1974) first, then turn to maximum likelihood computation to resolve ambiguities. Although important for small image processing systems, further computational improvements will become less and less cost effective as real time computational costs continue to fall through the development of fourth- and fifth-generation hardware computer systems.

Unsupervised methods rely on clustering measurement vectors according to some set of distance, similarity, or dispersion criteria. Many clustering heuristics have been devised and applied in image processing. Dubes and Jain (1976) provide a review and comparative analysis of a number of techniques which are commonly applied in pattern recognition. However, as Kendall (1972, p. 291) points out, clustering is a subjective matter to which little probabilistic theory is applicable. No clustering algorithms have as yet come to the fore which can incorporate prior knowledge in a formal fashion (except for the use of *à priori* starting vectors in interactive clustering) with an expected increment in class identification accuracy produced by the use of this additional information. However, recent developments involving guided clustering and automated labeling of unsupervised clusters blur the distinction between supervised and unsupervised techniques. Future work may well produce a continuum of intergrading methods from which a user can select a mix appropriate to the spatial, spectral, and temporal resolution of the data in hand and information output desired.

The purpose of this paper is to show how the use of prior information about the expected distribution of classes in a final classification map can be used in several different models to improve classification accuracies. Prior information is incorporated through the use of prior probabilities — that is, probabilities of occurrence of classes which are based on separate, independent knowledge concerning the area to be classified. Used in their simplest form, the probabilities weight the classes according to their expected distribution in the output dataset by shifting decision space boundaries to produce larger volumes in measurement space for classes which are expected to be large and smaller volumes for classes expected to be small.

The incorporation of prior probabilities into the maximum likelihood decision rule can also provide a mechanism for merging continuously measured observations (mul-

tispectral signatures) with discretely measured collateral variables such as rock type or soil type. As an example, consider an area of natural vegetation underlain by two distinctive rock types, each of which exhibits a unique mix of vegetation classes. Two sets of prior probabilities can be devised, one for each rock type, and the classifier can be modified to use the appropriate set of prior probabilities contingent on the underlying rock type. In this way, the classification process can incorporate discrete collateral information into the decision rule through a model contingent on an external conditioning variable. The method can also be extended to include two or more such discrete collateral datasets; the number is limited only by the ability to estimate the required sets of prior probabilities. Thus, prior probabilities provide a powerful mechanism for merging collateral datasets with multispectral images for classification purposes.

Another application of prior probabilities contingent upon a collateral dataset allows temporal weighting in a time-sequential classification system. As an example, consider distinguishing between two crop types which, through differing phenologies, can be easily separated early in the season but are confused later on in the growing period. Through the use of prior probabilities, a mid-summer classification can "look backward" to a spring classification to resolve ambiguity. Thus, winter wheat could be separated from spring wheat at mid-season by its distinctive early spring signature. This use of temporal information provides an alternative to the calculation of transformed vegetation indexes (TVI's) and comparable procedures (Richardson and Wiegand, 1977) in the identification of crops with multitemporal images (Rouse, et al., 1973). Such a time-sequential classification system could also be used to monitor land use change. In this case, a Markov-type predictive model is used directly to set prior probabilities based on patterns of change shown in an area.

Review of Maximum Likelihood Classification

To understand the application of prior probabilities to a classification problem, we must first review briefly the mathematics of the maximum likelihood decision rule. For the multivariate case, we assume each observation **X** (pixel) consists of a set of measurements on p variables (channels). Through some external procedure, we identify a set of observations which correspond to a class — that is, a set of similar objects characterized by a vector of means on measurement variables and a variance-covariance matrix describing the interrelationships among the measurement variables which are characteristic of the class. Although the parametric mean vector and dispersion matrix for the class remain unknown, they are estimated by the sample means and dispersion matrix associated with the object sample.

Multivariate normal statistical theory describes the probability that an observation **X** will occur, given that it belongs to a class k, as the following function:

$$\Phi_k(\mathbf{X}_i) = (2\pi)^{-\frac{1}{2}p}|\Sigma_k|^{-\frac{1}{2}}\, e^{(\mathbf{X}-\mu_k)'\Sigma_k^{-1}(\mathbf{X}-\mu_k)}. \tag{1}$$

(Table 1 presents an explanation of the symbols used in this and other expressions.) The quadratic product

$$\chi^2 = (\mathbf{X}-\mu_k)'\Sigma_k^{-1}(\mathbf{X}-\mu_k) \tag{2}$$

can be thought of as a squared distance function which measures the distance between the observation and the class mean as scaled and corrected for variance and covariance of the class. It can be shown that this expression is a χ^2 variate with p degrees of freedom (Tatsuoka, 1971).

As applied in a maximum likelihood decision rule, expression (1) allows the calculation of the probability that an observation is a member of each of k classes. The individual is then assigned to the class for which the probability value is greatest. In an opera-

tional context, we substitute observed means, variances, and covariances and use the log form of expression (1)

$$\ln[\Phi_k(\mathbf{X}_i)] = -\tfrac{1}{2}p\ln(2\pi) - \tfrac{1}{2}\ln|\mathbf{\Sigma}_k| - \tfrac{1}{2}(\mathbf{X}_i - \mathbf{m}_k)'\mathbf{D}_k^{-1}(\mathbf{X}_i - \mathbf{m}_k). \tag{3}$$

Since the log of the probability is a monotonic increasing function of the probability, the decision can be made by comparing values for each class as calculated from the right hand side of this equation. This is the decision rule that is used in the currently distributed versions of LARSYS and VICAR, two image processing program systems authored respectively by the Laboratory for Applications of Remote Sensing at Purdue University and the Jet Propulsion Laboratory of California Institute of Technology at Pasadena. A simpler decision rule, R_1, can be derived from expression (3) by eliminating the constants (Tatsuoka, 1971):

$$\begin{aligned} &R_1\text{: Choose } k \text{ which minimizes} \\ &F_{1,k}(\mathbf{X}_i) = \ln|\mathbf{D}_k| + (\mathbf{X}_i - \mathbf{m}_k)'\mathbf{D}_k^{-1}(\mathbf{X}_i - \mathbf{m}_k). \end{aligned} \tag{4}$$

The Use of Prior Probabilities in the Decision Rule

The maximum likelihood decision rule can be modified easily to take into account prior probabilities which describe how likely a class is to occur in the population of observations as a whole. The prior probability itself is simply an estimate of the proportion of the objects which will fall into a particular class. These prior probabilities are sometimes termed "weights," since the modified classification rule will tend to weigh more heavily those classes with higher prior probabilities.

Prior probabilities are incorporated into the classification through manipulation of the Law of Conditional Probability. To begin, we define two probabilities: $P\{\omega_k\}$, the probability that an observation will be drawn from class ω_k; and $P\{\mathbf{X}_i\}$, the probability of occurrence of the measurement vector $\mathbf{X}_i$. The Law of Conditional Probability states that:

$$P\{\omega_k|\mathbf{X}_i\} = \frac{P\{\omega_k, \mathbf{X}_i\}}{P\{\mathbf{X}_i\}}. \tag{5}$$

The probability on the left hand side of this expression will form the basis of a modified decision rule, since we wish to assign the ith observation to that class ω_k which has the highest probability of occurrence given the p-dimensional vector $\mathbf{X}_i$ which has been observed.

Again using the Law of Conditional Probability , we find that

$$P\{\mathbf{X}_i|\omega_k\} = \frac{P\{\omega_k, \mathbf{X}_i\}}{P\{\omega_k\}}. \tag{6}$$

In this expression, the left hand term describes the probability that the measurement vector will take on the values $\mathbf{X}_i$ given that the object measured is a member of class ω_k. This probability could be determined by sampling a population of measurement vectors for observations known to be from class ω_k; however, the distribution of such vectors is usually assumed to be Gaussian. Note that in some cases this assumption may not hold; as an example, Brooner et al. (1971) showed significantly higher classification accuracies for crops using simulated multispectral imagery with direct estimates of these conditional probabilities than with probabilities calculated according to Gaussian assumptions. However, the use of the multivariate normal approximation is widely accepted, and, in any case, it is only under rare circumstances that sufficient data are obtained to estimate the conditional probabilities directly.

Thus, we can assume that $P\{\mathbf{X}_i|\omega_k\}$ is acceptably estimated by $\Phi_k(\mathbf{X}_i)$ and rewrite expression (6) as

$$\Phi_k(\mathbf{X}_i) = \frac{P\{\omega_k, \mathbf{X}_i\}}{P\{\omega_k\}}. \tag{7}$$

Rearranging, we have

$$P\{\omega_k, X_i\} = \Phi_k(X_i) \cdot P\{\omega_k\} = \Phi_k^*(X_i). \tag{8}$$

Thus, we see that the numerator of expression (5) can be evaluated as the product of the multivariate density function $\Phi_k(X_i)$ and the prior probability of occurrence of class ω_k.

To evaluate the denominator of expression (5), we note that for all k classes the conditional probabilities must sum to 1:

$$\sum_{k=1}^{K} P\{\omega_k | X_i\} = 1 = \sum_{k=1}^{K} \left[\frac{\Phi_k(X_i) \cdot P\{\omega_k\}}{P\{X_i\}} \right]. \tag{9}$$

Therefore,

$$P\{X_i\} = \sum_{k=1}^{K} \Phi_k(X_i) \cdot P\{\omega_k\}. \tag{10}$$

Substituting (8) and (10) into (5),

$$P\{\omega_k | X_i\} = \frac{\Phi_k(X_i) \cdot P\{\omega_k\}}{\sum_{k=1}^{K} \Phi_k(X_i) \cdot P\{\omega_k\}} = \frac{\Phi_k^*(X_i)}{\sum_{k=1}^{K} \Phi_k^*(X_i)}. \tag{11}$$

The last expression, then, provides the basis for the decision rule which includes prior probabilities. Since the denominator remains constant for all classes, the observation is simply assigned to the class for which $\Phi_k^*(X_i)$, the product of $\Phi_k(X_i)$ and $P\{\omega_k\}$, is a maximum. In its simplest form, this decision rule can be stated as:

R_2: Choose k which minimizes
$F_{2,k}(X_i) = \ln|D_k| + (X_i - m_k)' D_k^{-1} (X_i - m_k) - 2 \ln P\{\omega_k\}$. (12)

This form of the decision rule is usually attributed to Tatsuoka and Tiedeman (1954; Tatsuoka, 1971).

It is important to understand how this decision rule behaves with different prior probabilities. If the prior probability $P\{\omega_k\}$ is very small, then its natural logarithm will be a large negative number; when multiplied by -2, it will become a large positive number and

thus $F_{2,k}$ for such a class will never be minimal. Therefore, setting a very small prior probability will effectively remove a class from the output classification. Note that this effect will occur even if the observation vector $\mathbf{X}_i$ is coincident with class mean vector $\mathbf{m}_k$. In such a case, the quadratic product distance function $(\mathbf{X}_i-\mathbf{m}_k)'\mathbf{D}_k^{-1}(\mathbf{X}_i-\mathbf{m}_k)$ goes to zero, but the prior probability term $-2\ln P\{\omega_k\}$ can still be large. Thus, it is entirely possible that the observation will be classified into a different class, one for which the distance function is quite large.

As the prior probability $P\{\omega_k\}$ becomes large and approaches 1, its logarithm will go to zero and $F_{2,k}$ will approach $F_{1,k}$ for that class. Since this probability and all others must sum to one, however, the prior probabilities of the remaining classes will be small numbers and their values of $F_{2,i}$ will be greatly augmented. The effect will be to force classification into the class with high probability. Therefore, the more extreme are the values of the prior probabilities, the less important are the actual observation values $\mathbf{X}_i$. This point is discussed in more detail in a following section.

Numerical Example

A simple numerical example may clarify this modification of the maximum likelihood decision rule. For this example, we assume two classes ω_1 and ω_2 in a two-dimensional measurement space. Their means and dispersion matrices are shown below.

$$\mathbf{m}_1 = [4\ \ 2] \qquad \mathbf{m}_2 = [3\ \ 3]$$
$$\mathbf{D}_1 = \begin{bmatrix} 3 & 4 \\ 4 & 6 \end{bmatrix} \tag{13}$$
$$\mathbf{D}_2 = \begin{bmatrix} 4 & 5 \\ 5 & 7 \end{bmatrix}$$

The determinants and inverses of these matrices are:

$$|D_1| = 2 \qquad |D_2| = 3$$

$$D_1^{-1} = \begin{bmatrix} 3 & -2 \\ -2 & \frac{3}{2} \end{bmatrix} \tag{14}$$

$$D_2^{-1} = \begin{bmatrix} \frac{7}{3} & -\frac{5}{3} \\ -\frac{5}{3} & \frac{4}{3} \end{bmatrix}$$

For this example, we wish to decide to which class the measurement vector (4,3) belongs.

To evaluate the probability associated with ω_1, we first evaluate the quadratic product

$$\chi^2 = (X-m_1)'D_1^{-1}(X-m_1) \tag{15}$$

$$\chi_1^2 = [0 \quad 1]\cdot\begin{bmatrix} 3 & -2 \\ -2 & \frac{3}{2} \end{bmatrix}\cdot\begin{bmatrix} 0 \\ 1 \end{bmatrix} = \frac{3}{2}. \tag{16}$$

The probability density value is then

$$\Phi_1(X) = \frac{1}{2\pi}\cdot\frac{1}{\sqrt{2}}\cdot e^{-\frac{1}{2}\cdot\frac{3}{2}} = .0532. \tag{17}$$

Similarly, for the second class,

$$\chi_2^2 = \begin{bmatrix} -1 & 0 \end{bmatrix}\cdot\begin{bmatrix} \frac{7}{3} & -\frac{5}{3} \\ -\frac{5}{3} & \frac{4}{3} \end{bmatrix}\cdot\begin{bmatrix} -1 \\ 0 \end{bmatrix} = 2 \tag{18}$$

$$\Phi_2(X) = \frac{1}{2\pi}\cdot\frac{1}{\sqrt{3}}\cdot e^{-\frac{1}{2}\cdot 2} = .0338 \tag{19}$$

Thus, the measurement vector (4,3) has a higher probability associated with membership in class ω_1 than in class ω_2, and it would be appropriate to classify the measurement into ω_1.

This same decision can be made by using the somewhat simpler decision rule R_1 (expression (4)):

$$F_{1,1}(X) = \ln|D_1| + (X - m_1)' D_1^{-1} (X - m_1) \tag{20}$$

$$F_{1,1}(X) = \ln(2) + \frac{3}{2} = 2.193; \tag{21}$$

$$F_{1,2}(X) = \ln|D_2| + (X - m_2)' D_2^{-1} (X - m_2) \tag{22}$$

$$F_{1,2}(X) = \ln(3) + 2 = 3.098. \tag{23}$$

Here again the decision is made to classify the observation X into ω_1.

The foregoing calculations assume equal probability of membership in ω_1 and ω_2. Removing this restriction, we take prior probabilities into account. Assume that the following prior probabilities are observed:

$$P\{\omega_1\} = \frac{1}{3} \qquad P\{\omega_2\} = \frac{2}{3} \tag{24}$$

Recalling the notation from expression (11) that $\Phi_k^*(X_i)$ denotes the probability density function adjusted for the prior probability,

$$\Phi_k^*(X_i) = \Phi_k(X_i) \cdot P\{\omega_k\}, \tag{25}$$

we calculate for the two classes

$$\Phi_1^*(X) = \Phi_1(X) \cdot P\{\omega_1\} = (.0532) \cdot \frac{1}{3} = .0177 \tag{26}$$

$$\Phi_2^*(X) = \Phi_2(X) \cdot P\{\omega_2\} = (.0338) \cdot \frac{2}{3} = .0225 \tag{27}$$

The actual conditional probabilities (expression (11)) then become

$$P\{\omega_1 | X_i = (4,2)\} = \frac{.0177}{.0177 + .0225} = .440 \tag{28}$$

$$P\{\omega_2|X_i=(4,2)\} = \frac{.0225}{.0177+.0225} = .560. \tag{29}$$

Thus, the prior probabilities modify the outcome of the decision rule, favoring class ω_2 for the observation (4,3) over class ω_1.

In terms of the decision rule R_2, we calculate

$$F_{2,1}(X) = \ln|D_1| + (X-m_1)'D_1^{-1}(X-m_1) - 2\ln P\{\omega_1\} \tag{30}$$

$$F_{2,1}(X) = F_{1,1}(X) - 2\ln P\{\omega_1\} \tag{31}$$

$$F_{2,1}(X) = 2.193 - 2\ln(\frac{1}{3}) = 4.390. \tag{32}$$

For the second class, the outcome is

$$F_{2,2}(X) = 3.098 - 2\ln(\frac{2}{3}) = 3.908. \tag{33}$$

Since the observation is classified into the class which minimizes the value of R_2, once again the second class is chosen.

Prior Probabilities Contingent on a Single External Conditioning Variable

Having shown how to modify the decision rule to take into account a set of prior probabilities, it is only a small step to consider several sets of probabilities, in which an external information source identifies which set is to be used in the decision rule. As an example, consider the effect of soil type on the distribution of crops that are likely to be grown in an area. In such a case, a single suite of crops will characterize the entire area, but the expected distribution of crops from one soil type to the next could be expected to vary considerably. Under these circumstances, it would be possible to collect a stratified random sample of the area to be classified, in order to quantify two sets of prior probabili-

ties: one for the crops on the first soil type, the other for crops on the second.

Thus, we introduce a third variable ν_j, which indicates the state of the external conditioning variable (e.g. soil type) associated with the observation. We wish, then, to find an expression describing

$$\mathbf{P}\{\omega_k|\mathbf{X}_i,\nu_j\}, \tag{34}$$

the probability that an observation will be a member of the class ω_k given its vector of observed measurements and the fact that it belongs to class ν_j of the external conditioning variable.

In deriving an expression to find this probability, we can make the assumption that the mean vector and dispersion matrix of the class will be the same regardless of the state of the external conditioning variable. This assumption is discussed more fully in a later section. The assumption implies that

$$\mathbf{P}\{\mathbf{X}_i|\omega_k\} = \mathbf{P}\{\mathbf{X}_i|\omega_k,\nu_j\}. \tag{35}$$

Expanding both sides of this relationship using the Law of Conditional Probability,

$$\frac{\mathbf{P}\{\mathbf{X}_i,\omega_k\}}{\mathbf{P}\{\omega_k\}} = \frac{\mathbf{P}\{\mathbf{X}_i,\omega_k,\nu_j\}}{\mathbf{P}\{\omega_k,\nu_j\}}. \tag{36}$$

Solving for the 3-way joint probability,

$$\mathbf{P}\{\mathbf{X}_i,\omega_k,\nu_j\} = \frac{\mathbf{P}\{\mathbf{X}_i,\omega_k\}\cdot\mathbf{P}\{\omega_k,\nu_j\}}{\mathbf{P}\{\omega_k\}}. \tag{37}$$

Substituting expression (8) into the left hand term of the right denominator,

$$\mathbf{P}\{\mathbf{X}_i,\omega_k,\nu_j\} = \frac{\Phi_k(\mathbf{X}_i)\cdot\mathbf{P}\{\omega_k\}\cdot\mathbf{P}\{\omega_k,\nu_j\}}{\mathbf{P}\{\omega_k\}} \tag{38}$$

$$\mathbf{P}\{\mathbf{X}_i,\omega_k,\nu_j\} = \Phi_k(\mathbf{X}_i)\cdot\mathbf{P}\{\omega_k,\nu_j\} = \Phi_k^{**}(\mathbf{X}_i). \tag{39}$$

Expanding expression (34) according to the Law of Conditional Probability,

$$P\{\omega_k|X_i,\nu_j\} = \frac{P\{\omega_k, X_i,\nu_j\}}{P\{X_i,\nu_j\}}. \tag{40}$$

Noting that since all classes are included, expression (40), when summed over all classes, must equal 1,

$$\sum_{k=1}^{K} P\{\omega_k|X_i,\nu_j\} = 1 = \frac{\sum_{k=1}^{K} P\{\omega_k, X_i,\nu_j\}}{P\{X_i,\nu_j\}}. \tag{41}$$

Rearranging,

$$P\{X_i,\nu_j\} = \sum_{k=1}^{K} P\{\omega_k, X_i,\nu_j\}. \tag{42}$$

Substituting expression (39) and (42) into (40), we have

$$P\{\omega_k|X_i,\nu_j\} = \frac{\Phi_k(X_i)\cdot P\{\omega_k,\nu_j\}}{\sum_{k=1}^{K}\Phi_k(X_i)\cdot P\{\omega_k,\nu_j\}} = \frac{\Phi_k^{**}(X_i)}{\sum_{k=1}^{K}\Phi_k^{**}(X_i)}. \tag{43}$$

This result is analogous to expression (11); note that the denominator remains constant for all k, and need not actually be calculated to select the class ω_k for which $\Phi_k^{**}(X_i)$ is a maximum.

The application of this expression in classification requires that the joint probabilities $P\{\omega_k,\nu_j\}$ be known. However, a simpler form using conditional probabilities directly obtained from a stratified random sample can be obtained through the application of the Law of Conditional Probability:

$$P\{\omega_k,\nu_j\} = P\{\omega_k|\nu_j\}\cdot P\{\nu_j\}. \tag{44}$$

Since $P\{\nu_j\}$ cancels from the numerator and denominator after substitution, we have

$$P\{\omega_k|X_i,\nu_j\} = \frac{\Phi_k(X_i)\cdot P\{\omega_k|\nu_j\}}{\sum_{k=1}^{K}\Phi_k(X_i)\cdot P\{\omega_k|\nu_j\}}. \tag{45}$$

Thus, either the joint or conditional probabilities may be used in the decision rule:

R_3: Choose k which minimizes
$$F_{3,k}(X_i) = \ln|D_k| + (X_i-m_k)'D_k^{-1}(X_i-m_k) - 2\ln P\{\omega_k,\nu_j\} \tag{46}$$

R'_3: Choose k which minimizes
$$F'_{3,k}(X_i) = \ln|D_k| + (X_i-m_k)'D_k^{-1}(X_i-m_k) - 2\ln P\{\omega_k|\nu_j\}. \tag{47}$$

Numerical Example

To illustrate this use of prior probabilities contingent on an external conditioning variable, let us return to the two-class example discussed earlier. This time, however, let us assume that a stratified random sample of the area to be classified produces the estimates of probabilities shown in Table 2. The conditioning variable ν_j has two states: ν_1 and ν_2. Under the conditions of ν_1, both classes have equal prior probabilities; under ν_2, the second class is more likely to appear, with the probability of .7 for ω_2 and .3 for ω_1. Table 3 presents the calculations for this example. For ν_1, ω_1 would be the most likely choice. In the case of ν_2, ω_2 is more probable.

Adding Additional Conditioning Variables

Logic analogous to that of the preceding section shows that classification decisions may be made contingent on any number of external multistate conditioning variables. However, a separate set of prior probabilities must be estimated for all possible states of conditioning variables. For example, consider classifying natural vegetation in an area containing four distinctive rock types, six different soil types, and four unique topographic habitats. Ninety-six sets of prior probabilities will then be required. Estimating these pro-

babilities by separate samples would be prohibitive for such a large number of combinations.

To alleviate this problem, it is possible to model these probabilities from a much smaller set under the assumption of no high-level interaction. This procedure amounts to the calculation of expected values for a multidimensional contingency table when only certain marginal totals are known. Techniques for such modeling have been described in the recent statistical literature, and are summarized in two current books by Bishop, Fienburg and Holland (1975) and by Upton (1978). (Other treatments appear in Cox (1970) and Fienburg (1977).) The discussion below is based partly on the treatment presented in pp. 57-101 of Bishop, et al., and the reader is referred to these works for cases involving modeling beyond the trivariate case presented here.

As a simple example, consider the three-way case in which a measurement is a member of class ω_k and is also associated with two conditioning variables ν_j and o_l. Then the Law of Conditional Probability states

$$\mathbf{P}\{\omega_k|\nu_j,o_l\} = \frac{\mathbf{P}\{\omega_k,\nu_j,o_l\}}{\sum_{k=1}^{K}\mathbf{P}\{\omega_k,\nu_j,o_l\}}, \tag{48}$$

since $\mathbf{P}\{\nu_j,o_l\} = \sum_{k=1}^{K}\mathbf{P}\{\omega_k,\nu_j,o_l\}$. There are $K\times J\times L$ probabilities of the form $\mathbf{P}\{\omega_k,\nu_j,o_l\}$, and we wish to estimate these with maximum likelihood without sampling the full set. Such an estimate is possible, assuming no three-way interaction between ω,ν, and o, if probabilities of the forms $\mathbf{P}\{\omega_k,\nu_j\}$, $\mathbf{P}\{\omega_k,o_l\}$, and $\mathbf{P}\{\nu_j,o_l\}$ are known.

The method, first described by Deming and Stephan (1940), requires iterative fitting of three-way probabilities $\mathbf{P}\{\omega_k,\nu_j,o_l\}$ to conform with observed two-way probabilities $\mathbf{P}\{\omega_k,\nu_j\}$, $\mathbf{P}\{\omega_k,o_l\}$, and $\mathbf{P}\{\nu_j,o_l\}$. Beginning with an initial starting probability $\hat{\mathbf{P}}_0\{\omega_k,\nu_j,o_l\}$, the individual three-way probabilities are first rescaled to conform with one

set of two-way probabilities:

$$\hat{\mathbf{P}}_1\{\omega_k,\nu_j,o_l\} = \hat{\mathbf{P}}_0\{\omega_k,\nu_j,o_l\}\cdot\frac{\sum_{l=1}^{L}\hat{\mathbf{P}}_0\{\omega_k,\nu_j,o_l\}}{\mathbf{P}\{\nu_j,o_l\}} \tag{49}$$

Rescaling then proceeds for another set of two-way conditionals,

$$\hat{\mathbf{P}}_2\{\omega_k,\nu_j,o_l\} = \hat{\mathbf{P}}_1\{\omega_k,\nu_j,o_l\}\cdot\frac{\sum_{j=1}^{J}\hat{\mathbf{P}}_1\{\omega_k,\nu_j,o_l\}}{\mathbf{P}\{\omega_k,o_l\}}, \tag{50}$$

and finally for the last set:

$$\hat{\mathbf{P}}_3\{\omega_k,\nu_j,o_l\} = \hat{\mathbf{P}}_2\{\omega_k,\nu_j,o_l\}\cdot\frac{\sum_{l=1}^{L}\hat{\mathbf{P}}_2\{\omega_k,\nu_j,o_l\}}{\mathbf{P}\{\omega_k,\nu_j\}}. \tag{51}$$

As the procedure is repeated, convergence occurs rapidly, and values stabilize within a small number of iterations (Deming and Stephan, 1940). The method always converges toward the unique set of maximum likelihood estimates and can be used with any set of starting values; further, estimates may be determined to any preset level of accuracy (Bishop, Fienburg, and Holland, 1975, p. 83).

In a typical remote sensing application, a stratified random sample is collected which estimates the two conditional probability sets $\mathbf{P}\{\omega_k|\nu_j\}$ and $\mathbf{P}\{\omega_k|o_l\}$. In addition, probabilities of the form $\mathbf{P}\{\omega_k,o_l\}$ are obtained by processing registered digital images of maps showing the spatial distributions of ν and o. By noting that

$$\mathbf{P}\{\nu_j\} = \sum_{l=1}^{L}\mathbf{P}\{\nu_j,o_l\} \tag{52}$$

and using the Law of Conditional Probability,

$$\mathbf{P}\{\omega_k,\nu_j\} = \mathbf{P}\{\omega_k|\nu_j\}\cdot\mathbf{P}\{\nu_j\}, \tag{53}$$

the joint probabilities $\mathbf{P}\{\omega_k,\nu_j\}$ and $\mathbf{P}\{\omega_k,o_l\}$ can easily be calculated from the sets of

conditional probabilities $\mathbf{P}\{\omega_k|\nu_j\}$ and $\mathbf{P}\{\omega_k|o_l\}$.

Numerical Example

A simple numerical example will illustrate the iterative method. Table 4 presents a set of one-way conditional probabilities for an example of three classes ω_k, k=1,2,3 with two conditioning variables ν_j, j=1,2,3 and o_l, l=1,2,3. Although simple decimal values are assumed here for ease of computation, these values would normally be obtained by prior random stratified sampling. Also required are the joint probabilities $\mathbf{P}\{\nu_j, o_l\}$ (Table 5). Tables 6 and 7 show how values for $\mathbf{P}\{\omega_k, \nu_j\}$ and $\mathbf{P}\{\omega_k, o_l\}$ are calculated according to expressions (52) and (53).

Table 8 presents the results of the first two iterations in fitting the no-two-way-interaction models to these data. Using the criterion of no further change in any $\mathbf{P}\{\omega_k, \nu_j, o_l\}$ of greater than 10^{-6}, convergence is reached at iteration 23. Although these probabilities can be used directly in decision rule R_3, it may be easier to examine the values as conditional probabilities as used in R'_3. These values are shown in the last column of the table.

Time-Sequential Classification

If a classification carried out at a earlier time is viewed as an external conditioning variable, then the mechanism of prior probabilities can be used to make the outcome of a classification contingent on the earlier classification. This application is best clarified by an example. Consider an agricultural classification with four field types: rice, cotton, orchard, and fallow. An early spring classification reveals the presence of young rice with high accuracy, but at that time cotton cannot be distinguished from fallow fields. Orchards are easily distinguished from field crops at any time of year. By early summer, many fields which classified as fallow are likely to be in cotton; however, fields classified as rice are

still likely to be rice. Orchards will remain unchanged in areal extent.

Data from prior years are collected to quantify these expected changes, and a transition probability matrix is devised which describes the changes in classification expected during the early spring-early summer period (Table 9). In this example, spring classification shows thirty percent of the observations to be rice; by summer, ninety percent of these observations are expected to continue as rice, with ten percent returning to fallow because of crop failure or lack of irrigation. Twenty percent of the spring observations are orchards, and all of these are expected to remain in orchard through early summer. Fallow fields, constituting fifty percent of the spring observations, are most likely to become cotton (probability .7), with a few becoming rice, orchard, or remaining fallow (probabilities .1). Since no observations are classified as cotton in spring, no transition probabilities are needed for that class. This use of transition probabilities was suggested as early as 1967 by Simonett, et al.

The transition probability matrix can also be recognized as a matrix of conditional probabilities $\mathbf{P}\{\omega_k|\nu_j\}$ which describe the probability that an observation will fall into summer class ω_k given that the observation falls into spring class ν_j. Thus, the early summer classification can be made contingent on the early spring classification through the prior probability mechanisms discussed earlier, and any possible confusion between cotton and rice in summer will be resolved by the spring classification. It is also interesting to note that the transition probability matrix is actually a square stochastic matrix, and therefore the situation is equivalent to a simple, one-step Markov process. Under these conditions, the expected posterior probabilities $\mathbf{P}\{\omega_k\}$ are

$$\mathbf{P}\{\omega_k\} = \sum_{j=1}^{J} \mathbf{P}\{\omega_k|\nu_j\} \cdot \mathbf{P}\{\nu_j\}. \tag{54}$$

In a recent paper, Swain (1978) has carried this approach a step further, incorporating in the decision rule both measurement vectors $\mathbf{X}_{i,1}$ and $\mathbf{X}_{i,2}$ taken from times t_1 and t_2 at point i in space. In contrast, the approach described above uses $\mathbf{X}_{i,1}$ to predict ν_j at time t_1 and then uses $\mathbf{X}_{i,2}$ and ν_j to make the classification decision at time t_2. Swain's decision rule, in the notation of this paper, becomes

R_S: Choose k to maximize

$$F_{S,k}(\mathbf{X}_{i,1}, \mathbf{X}_{i,2}) = \sum_{j=1}^{J} P\{\mathbf{X}_{i,1}|\nu_j\} \cdot P\{\mathbf{X}_{i,2}|\omega_k\} \cdot P\{\omega_k|\nu_j\} \cdot P\{\nu_j\}. \tag{55}$$

Swain has termed this rule the "cascade classifier."

Swain's approach has the advantage of using full information about the distances of the measurement vectors $\mathbf{X}_{i,1}$ and $\mathbf{X}_{i,2}$ from class means; however, as Swain notes, there is no way to make the first observation set dominate the second. When the transition probability matrix goes to an identity matrix, the classification rule becomes:

$$F_{S,k}(\mathbf{X}_{i,1}, \mathbf{X}_{i,2}) = P\{\mathbf{X}_{i,1}|\nu_j\} \cdot P\{\mathbf{X}_{i,2}|\omega_k\} \cdot P\{\nu_j\}, \tag{56}$$

and the two observations become equally weighted. Decision rule R_3 does allow the first observation to dominate; here, an identity transition probability matrix will preserve the first classification completely. On the other hand, R_3 assumes that the prior classification is perfectly correct, and any errors in the prior classification will also be preserved to an extent controlled by the transition probabilities. Thus, both approaches are relevant, depending on the classification task at hand.

Remote Sensing Example

The preceding numerical examples have demonstrated the application of prior probabilities in maximum likelihood classification in a computational context; a real example drawn from remote sensing will hopefully serve to further understanding in an operational

context. This example (Strahler, Logan, and Bryant, 1978) is drawn from a problem involving classification of natural vegetation in a heavily forested area of northern California. In the classification, spectral data are used to define species-specific timber types, and elevation and slope aspect are used as collateral data channels to improve classification accuracy.

The area selected for application of the classification techniques described above is referred to as the Doggett Creek study area, comprising about 220 sq. km. of private and publicly-owned forest land in northern California near the town of Klamath River. Located within the Siskiyou Mountains, elevations in the area range from 500 m at the Klamath River, which crosses the southern portion, to 2065 m near Dry Lake Lookout on an unnamed summit. A well developed network of logging roads and trails is present, providing relatively easy access to nearly all of the area by road or foot.

A wide variety of distinctive vegetation types is present in the area. Life-form classes include alpine meadow, fir park, pasture, cropland, and burned, reforested areas. Forest vegetation includes, from high elevation to low elevation, such types as red fir, white fir, douglas fir-ponderosa pine-incense cedar, pine-oak, and oak-chapparal. Thus, the topographic and vegetational characteristics of the area are well differentiated.

After a review of available Landsat frames which included the Doggett Creek area, two were selected for analysis: July 4, 1973, and October 15, 1974. The two frames were obtained as computer compatible tapes from the EROS Data Center, Sioux Falls, S.D. and then reformatted and precision rectified to sinusoidal projections. Pixel size was converted to 80 x 80 meters in the rectification process to facilitate film writer playback. Using the July image as a base, the October frame was registered to within a half-pixel error using seven control points. In this process, the October image was resampled to conform with the July image using a cubic spline convolution algorithm. Figure 1

presents an image of the study area using Landsat band 5 (.6 - .7 μ) from the July frame.

Also registered to the July image was a terrain image derived from the U.S. Geological Survey 1:250,000 digital terrain tape for the Weed, California, quadrangle. In the registration process, the image was converted to 80 x 80 meter pixel size, and stretched to yield a full range of gray tone values. Slope and aspect images were generated directly from the registered elevation data using a least squares algorithm which fits a plane to each pixel and its four nearest neighbors.

The slope aspect image consisted initially of gray tone densities between 0 (black) and 255 (white) which indicated the azimuth of slope orientation, ranging clockwise from 0° to 359°. These values were then transformed by a cosine function proposed by Hartung and Lloyd (1969). Since northeast slopes present the most favorable growing environment, and southwest slopes the least favorable, with northwest and southeast slopes of neutral character, the density tones of azimuths were rescaled with 3 representing due southwest and 255 representing due northeast. (Values of 0, 1, and 2 were reserved for special codings.) Neutral slopes, oriented northwest or southeast, thus received density tones near 127. The function also corrected automatically for the 12° skew of the Landsat image. For processing as collateral data channels, both elevation and transformed aspect were converted to three-state variables: elevation to low, middle, and high; and aspect to southwest, neutral (southeast or northwest) and northeast. Figure 2 shows elevation and aspect images as well as their three-state versions.

Following an initial reconnaissance of the area, thirteen species-specific forest cover classes were selected as representing the range of cover types within the study area. These classes were defined by a set of 93 training sites ranging in size from approximately twenty to one hundred pixels. Further processing revealed the presence of several subtypes within most of the forest cover classes. For example, open canopy douglas fir train-

ing sites were divided into two subtypes. Such subtypes were also defined for hardwood, white fir, douglas fir, sparce, and grass and shrub cover classes. Throughout the classification procedure, these subtypes were kept separate, joining together only in the final classification map.

In order to obtain estimates of prior probabilities for the forest cover class types, one hundred points were randomly selected from a grid covering the Doggett Creek study area by drawing coordinates from a random number table. At each of these points, the forest cover class was determined either by interpretation of 1:8,000 color aerial photography, or by actual field visit. Of the 100 points, 15 were discarded because they fell (1) in locations which were inaccessable in the time available; or (2) outside the area covered by 1:8,000 air photos (and therefore could not be accurately located on either the Landsat frame or on the ground). At each point, the elevation and aspect class was also recorded, thus allowing type counts to be cross tabulated according to elevation and aspect.

From this sample of 85 points, three sets of probabilities were prepared. The first of these recorded the unconditional prior probabilities of the forest cover types — that is, their proportional representation within the entire study area. The second and third sets of probabilities aggregated the points according to elevation and aspect classes, and were used to estimate three sets of probabilities for each topographic parameter (low, middle, and high for elevation, and northeast, neutral and southwest for aspect). Table 10 shows how the classes were defined, and describes the number of points falling into each.

These estimates of probabilities lack precision because the number of sample points is small; with 85 samples distributed across 13 cover types, the calculated probabilities are more likely to indicate adequately the rank order of the magnitudes rather than the true values of the magnitudes themselves. However, under constraints of field time and expense, it was not possible to prepare a larger dataset for this particular trial. Consider-

ing the sensitivity of the classification to extreme probability values, future work should estimate these probability sets to a higher degree of accuracy.

This dataset was also used to estimate classification accuracies. By recording the pixel location of each of the sample points, the cover type as determined on the ground could be compared with the cover type as classified on the Landsat image. Because of uncertainties in locating each pixel on the 1:8,000 air photos, it was necessary to specify alternate acceptable classifications for each point. For example, a pixel falling into a stand identified on the ground as douglas fir, open canopy might fall almost entirely on a clearing, and thus be classified as grass/shrub, or sparse if containing a few canopy trees. In such a case, the classification was termed correct. On the other hand, classifications such as hardwood, alpine meadow, or red fir forest would be an obvious error in a douglas fir stand. Here again, estimated accuracies are influenced by the limited size of the sample.

Note that the field data are used to produce a classification which is then assessed for accuracy by reference to the same data. Separate samples would clearly be more desirable. The decision to use the same set of samples to determine accuracy that was used to estimate the probability sets was, again, influenced by available field time and travel funds. However, the accuracies are greatly dependent on the spatial location of the data points; only in the aggregate does each data point influence the classification. Thus, we would expect the accuracies to reflect only a slight positive bias produced by this cost-reducing strategy.

Although several different classifications were carried out, only three are of importance here. The first used spectral data only, and assumed equal prior probabilities; this classification yielded an accuracy of 58 percent (Figure 3). In the second, three sets of prior probabilities for the forest types were used, each contingent on one of the three elevation states (Table 10; Fig. 4). The classification software was modified to use a table

look-up of prior probabilities with elevation class as one index into the prior probability table. This technique increased classification accuracy from 58 percent to 71 percent.

The third classification used two sets of prior probabilities contingent on elevation and aspect, analogous to $P(\omega_k|v_j)$ and $P(\omega_k|o_l)$ in the preceding section. Software then systematically sampled the registered elevation class and terrain class images to yield the joint probabilities of elevation and aspect classes, analogous to $P(v_j, o_l)$. The iterative algorithm described earlier was then applied to estimate the set of conditional probabilities for forest cover classes contingent on all combinations of elevation and aspect classes. Classification using these estimated probabilities contingent on both elevation and aspect yielded an accuracy of 77 percent, an improvement over that observed for elevation alone (Figure 5). Thus, this example demonstrates how prior probabilities can be used to merge continuous variables of multispectral brightness with discrete variables of elevation and aspect class to improve classification accuracies.

Discussion

As noted earlier, extreme values of prior probabilities can force a classification to be made essentially without information concerning the observation itself. When priors are equal, however, they have no effect. The classifier, then, continuously trades off the role of the multivariate information for the role of prior information, depending on both the magnitude of the distance of the multivariate observation vector from the class mean vector and the ratios of the particular prior probabilities involved in the decision. When the experimental design allows the priors to be determined by external conditioning variables, the effect is to classify based on multivariate information when the possible classes are not particularly influenced by the conditioning variable and to classify based on prior information when multivariate data are equivocal or some classes are much more or less likely than others. Thus, in the forest classification example, terrain information served to

differentiate species-specific cover types (e.g., red fir, white fir), whereas spectral information differentiated life form classes (e.g., grass-shrub, hardwood).

Another important point concerns the dependence of the prior probabilities on the scene itself; the relative areas of the classes in the output scene must be accurately estimated. If the output scene shifts in area, then the priors must be changed. The classification cannot be extended to a new area without reestimating the prior probabilities. Thus, it would be appropriate to use county crop acreage values to set prior probabilities only when the entire area of the county, no more, no less, is to be classified. In the case of one or two external conditioning variables determining the appropriate set of priors, both the joint probabilities $\mathbf{P}\{\nu_j, o_l\}$ and the conditional probabilities $\mathbf{P}\{\omega_k|\nu_j\}$ and $\mathbf{P}\{\omega_k|o_l\}$ must truly represent the area to be classified, for, taken together, they determine the prior probability values actually used in computation. In some applications, it may be possible to extend the conditionals to a new scene in which only the joint probabilities of the variables change — for example, a forest cover classification with elevation and aspect as conditioning variables which is extended from one uniform area to another. The new area will have different joint probabilities $\mathbf{P}\{\nu_j, o_l\}$ (and derived priors $\mathbf{P}\{\nu_j\}$ and $\mathbf{P}\{o_l\}$), but it might be reasonable to assume that the conditional probabilities are ecologically based and remain consistent from one area to the next.

It should be noted that collateral information cannot be incorporated through the prior probability mechanism without the collection of data to estimate the priors and/or conditionals. If the collateral data are likely to be unrelated to the multivariate data and are expected to influence strongly the prior probabilities of the classes, then such estimating costs will be justified, for significant improvements in accuracy should result. Ultimately, it is up to the user to balance the costs of acquiring such information with the value of the expected payoffs in accuracy which are anticipated.

The logic which culminates in decision rules R_2 and R_3 assumes that the mean vector and dispersion matrix for a class are not affected by the external conditioning variable (see expression (35)) — in the remote sensing case, this means that the signatures are invariant. In some applications, this assumption may not be warranted. An example would be an agricultural crop classification with soil type as a collateral variable. Here the signature of the soil itself, at least in the earlier states of crop development, will influence the crop signature. In this situation, there is no recourse but to spectrally characterize each combination of crop and soil type so that probabilities of the form $P\{\omega_k|\mathbf{X}_i, \nu_j\}$ can be calculated. Following the logic of expressions (5) through (11), it is possible to show that

$$P\{\omega_k|\mathbf{X}_i, \nu_j\} = \frac{P\{\mathbf{X}_i|\omega_k, \nu_j\} \cdot P\{\omega_k|\nu_j\}}{\sum_{k=1}^{K} P\{\mathbf{X}_i|\omega_k, \nu_j\}}, \tag{57}$$

which could be made the basis of a decision rule related to R_2.

A final point worthy of discussion concerns modeling of joint probabilities, suggested in an earlier section to reduce the need for more extensive ground sampling. The model presented is but one example of a large variety of techniques by which collateral data can be used to predict the spatial distribution of classes in an output image. Discrete and continuous collateral variables can be merged either using empirical techniques including multiple regression, logit analysis, discriminant analysis, analysis of covariance, and contingency table analysis, or by constructing more functional models which model the spatial processes actually occurring in a deterministic way. When such models are constructed and interfaced with remotely sensed data, the result may be extremely powerful, both for the ability to accurately predict a spatial pattern and for the understanding of the complex system which produces it.

Conclusions

The use of prior probabilities in maximum likelihood classification allows:

(1) the incorporation into the classification of prior knowledge concerning the frequencies of output classes which are expected in the area to be classified;

(2) the merging of one or more discrete collateral datasets into the classification process through the use of multiple prior probability sets describing the expected class distribution for each combination of discrete collateral variables; and

(3) the use of time-sequential information in making the outcome of a later classification contingent on an earlier classification.

Thus, prior probabilities can be a powerful and effective aid to improving classification accuracy and modeling the behavior of spatial systems.

Acknowledgements

The research reported herein was supported in part by NASA grant NSG-2377, NASA contract NAS-9-15509, and the California Institute of Technology's President's Fund (award PF-123), under NASA contract NAS-7-100. I am greatly indebted to D. S. Simonett, P. H. Swain, R. M. Haralick, and W. Wigton for critical review of the manuscript.

References

Addington, J. D. (1975), *VICAR Program FASTCLAS,* VICAR Documentation, Image Processing Laboratory, Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA, 3 pp.

Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. (1975), *Discrete Multivariate Analysis: Theory and Practice,* MIT Press, Cambridge, MA, 557 pp.

Brooner, W. G., Haralick, R. M., and Dinstein, I. (1971), Spectral parameters affecting automated image interpretation using Bayesian probability techniques, *Proc. Seventh Int. Symp. on Remote Sens. of Environ.,* pp. 1929-1948.

Chow, C. K. (1957), An optimum character recognition system using decision functions, *IREE Trans. Electron. Computers* 6: 247-254.

Cox, D. R. (1970), *The Analysis of Binary Data,* Methuen & Co., Ltd., London, 142 pp.

Deming, W. E. and Stephan, F. F. (1940), On a least squares adjustment of a sampled frequency table when the expected marginal totals are known, *Ann. Math. Statis.* 11: 427-444.

Dubes, R. and Jain, A. K. (1976), Clustering techniques: The user's dilemma, *Pattern Recognition* 8: 247-260.

Fienberg, S. E. (1977), *The Analysis of Cross-Classified Categorical Data,* MIT Press, Cambridge, MA, 151 pp.

Goodenough, D., and Shlien, S. (1974), Results of cover-type classification by maximum likelihood and parallelepiped methods, *Proc. Second Canadian Symp. on Remote Sensing,* 1: 136-164.

Hartung, R. E., and Lloyd, W. J. (1969), Influence of aspect on forests of the Clarksville Soils In Dent County, Missouri, *J. Forestry* 67: 178-182.

Kendall, M. G. (1972), in *Frontiers of Pattern Recognition* (Satosi Watanabe, Ed.), Academic Press, New York, pp. 291-307.

Nilsson, N. J. (1965), *Learning Machines — Foundations of Trainable Pattern-Classifying Systems,* McGraw-Hill Book Co., New York.

Reeves, R. G., Anson, A., and Landen, D. (1975), *Manual of Remote Sensing,* Amer. Soc. of Photogrammetry, Falls Church, VA, 2 vols., 2144 pp.

Richardson, A. J. and Wiegand, C. L. (1977), Distinguishing vegetation from soil background information, *Photogrammetric Engr. and Remote Sensing* 43: 1541-1552.

Rouse, J. W., Jr., Hass, R. H., Schell, J. A., and Deering, D. W. (1973), Monitoring vegetation systems in the Great Plains with ERTS, *Third ERTS Symposium,* NASA Special Publication SP-351, 1:309-317.

Sebestyen, G. (1962), *Decision-Making Processes in Pattern Recognition,* Macmillan, New York.

Schell, J. A. (1973), in *Remote Sensing of Earth Resources, Volume I* (F. Shahrokhi, Ed.), University of Tennessee Space Institute, Tullahoma, TN, pp. 374-394.

Schlien, S., and Smith, A. (1975), A rapid method to generate spectral theme classification of Landsat imagery, *Remote Sens. Environ.* 4: 67-77.

Simonett, D. S., Eagleman, J. E., Erhart, A. B., Rhodes, D. C., and Schwarz, D. E. ((1967), *The Potential of Radar as a Remote Sensor in Agricultu · I. A Study with K-Band Imagery in Western Kansas,* CRES Report No. 61-21, University of Kansas, Lawrence, KN, 13 pp.

Strahler, A. H., Logan, T. L., and Bryant, N. A. (1978), Improving forest cover classification accuracy from Landsat by incorporating topographic information, *Proc. Twelfth Int. Symp. on Remote Sens. of Environ.,* pp. 927-942.

Swain, P. H. (1978), Bayesian classification in a time-varying environment, *IEEE Trans. on Systems, Man and Cybern.,* SMC-8: 879-883.

Tatsuoka, M. M.(1971), *Multivariate Analysis: Techniques for Educational and Psychological Research,* John Wiley & Sons, New York, 310 pp.

Tatsuoka, M. M. and Tiedeman, D. V. (1954), Discriminant Analysis, *Rev. of Ed. Res.* 25: 402-420.

Upton, G. J. G. (1978), *The Analysis of Cross-Tabulated Data,* John Wiley & Sons, New York, 148 pp.

Tables for Prior Probabilities

TABLE 1 Notation

TERM	DEFINITION
p	Number of measurement variables used to characterize each object or observation.
$\mathbf{X}$	A p-dimensional random vector.
$\mathbf{X}_i$	Vector of measurements on p variables associated with the ith object or observation; $i=1,2,\ldots,N$.
$\mathrm{P}\{\mathbf{X}_i\}$	Probability that a p-dimensional random vector $\mathbf{X}$ will take on observed values $\mathbf{X}_i$.
ω_k	Member of the kth set of classes ω; $k=1,2,\ldots,K$.
ν_j	Member of the jth set of states for a conditioning variable ν; $j=1,2,\ldots,J$.
$\mathrm{P}\{\omega_k\}$	Probability that an observation will be a member of class ω_k; prior probability for class ω_k.
$\mathrm{P}\{\nu_j\}$	Probability that an observation will be associated with state j of conditioning variable ν; prior probability for state ν_j.
$\mathrm{P}\{\omega_k\|\mathbf{X}_i\}$	Probability that an observation is a member of class ω_k given that measurement vector $\mathbf{X}_i$ is observed.
$\Phi_k(\mathbf{X}_i)$	Probability density value associated with observation vector $\mathbf{X}_i$ as evaluated for class k.
μ_k	Parametric mean vector associated with the kth class.
$\mathbf{m}_k$	Mean vector associated with a sample of observations belonging to the kth class; taken as an estimator of μ_k.
$\mathbf{\Sigma}_k$	Parametric p by p dispersion (variance-covariance) matrix associated with the kth class.
$\mathbf{D}_k$	p by p dispersion matrix associated with a sample of observations belonging to the kth class; taken as an estimator of $\mathbf{\Sigma}_k$.

TABLE 2 Simple Prior Probabilities for Numerical Example

	Conditioning Variable	
Probability	ν_1	ν_2
$P\{\omega_1\}$	.5	.3
$P\{\omega_2\}$	.5	.7

TABLE 3 Calculation of Maximum Likelihood Posterior Probabilities

	ν_1		ν_2	
	ω_1	ω_2	ω_1	ω_2
$\Phi_k(\mathbf{X}_i)$	.0532	.0338	.0532	.0338
$P\{\omega_k\|\nu_j\}$	.0266	.0169	.0160	.0237
$\Phi_k(\mathbf{X}_i)\cdot P\{\omega_k\|\nu_j\}$	.0266	.0169	.0169	.0237
$\sum_{k=1}^{2}\Phi_k(\mathbf{X}_i)\cdot P\{\omega_k\|\nu_j\}$	.0435		.0397	
$P\{\omega_k\|\mathbf{X}_i,\nu_j\}$	.611	.389	.403	.597

TABLE 4 Conditional Probabilities for Numerical Example

	$P\{\omega_k,\nu_j\}$			$P\{\omega_k,o_l\}$		
ω	ν_1	ν_3	ν_3	o_1	o_2	o_3
ω_1	.6	.5	.2	.5	.8	.2
ω_2	.3	.2	.4	.4	.1	.3
ω_3	.1	.3	.4	.1	.1	.5

TABLE 5 Joint Probabilities for Numerical Example

	$P\{\nu_j,o_l\}$			
ν	o_1	o_2	o_3	$P\{\nu_j\}$
ν_1	.08	.12	.14	.34
ν_2	.07	.09	.12	.28
ν_3	.16	.10	.12	.38
$P\{o_l\}$	.31	.31	.38	

TABLE 6 Calculation of Joint Two-Way Probabilities for Numerical Example

	$P[\omega_k \mid \nu_1]$		$P[\nu_1]$		$P[\omega_k, \nu_1]$	$P[\omega_k \mid \nu_2]$		$P[\nu_2]$		$P[\omega_k, \nu_2]$	$P[\omega_k \mid \nu_3]$		$P[\nu_3]$		$P[\omega_k, \nu_3]$
ω_1	.6	×	.34	=	.204	.5	×	.28	=	.140	.2	×	.38	=	.076
ω_2	.3	×	.34	=	.102	.2	×	.28	=	.056	.4	×	.38	=	.152
ω_3	.1	×	.34	=	.034	.3	×	.28	=	.084	.4	×	.38	=	.152

TABLE 7 Calculation of Joint Two-Way Probabilities for Numerical Example

	$P[\omega_k \mid o_1]$		$P[o_1]$		$P[\omega_k, o_1]$	$P[\omega_k \mid o_2]$		$P[o_2]$		$P[\omega_k, o_2]$	$P[\omega_k \mid o_3]$		$P[o_3]$		$P[\omega_k, o_3]$
ω_1	.5	×	.31	=	.155	.8	×	.31	=	.248	.2	×	.38	=	.076
ω_2	.4	×	.31	=	.124	.1	×	.31	=	.031	.3	×	.38	=	.114
ω_3	.1	×	.31	=	.031	.1	×	.31	=	.031	.5	×	.38	=	.190

TABLE 8 Iterative Fitting of No-Three-Way Interaction Model

			INITIAL	ITERATION 1			ITERATION 2			FINAL	
k	j	l	$\hat{P}_0[\omega_k, \nu_j, o_l]$	$\hat{P}_1[\omega_k, \nu_j, o_l]$	$\hat{P}_2[\omega_k, \nu_j, o_l]$	$\hat{P}_3[\omega_k, \nu_j, o_l]$	$\hat{P}_1[\omega_k, \nu_j, o_l]$	$\hat{P}_2[\omega_k, \nu_j, o_l]$	$\hat{P}_3[\omega_k, \nu_j, o_l]$	$\hat{P}[\omega_k, \nu_j, o_l]$	$\hat{P}[\omega_k \mid \nu_j, o_l]$
1	1	1	.0370	.0680	.0753	.0502	.0487	.0635	.0557	.0614	.7674
1	1	2	.0370	.0680	.1205	.1074	.1043	.1196	.1098	.1132	.9431
1	1	3	.0370	.0680	.0369	.0525	.0510	.0457	.0494	.0456	.3546
1	2	1	.0370	.0467	.0517	.0432	.0408	.0532	.0479	.0517	.7383
1	2	2	.0370	.0467	.0826	.0760	.0718	.0823	.0785	.0820	.9116
1	2	3	.0370	.0467	.0253	.0289	.0274	.0245	.0250	.0221	.1845
1	3	1	.0370	.0253	.0280	.0422	.0293	.0382	.0434	.0386	.2413
1	3	2	.0370	.0253	.0449	.0579	.0402	.0461	.0542	.0504	.5038
1	3	3	.0370	.0253	.0138	.0093	.0065	.0058	.0052	.0023	.0196
2	1	1	.0370	.0340	.0408	.0272	.0309	.0262	.0230	.0179	.2243
2	1	2	.0370	.0340	.0102	.0091	.0103	.0088	.0081	.0059	.0492
2	1	3	.0370	.0340	.0375	.0534	.0607	.0544	.0589	.0613	.4376
2	2	1	.0370	.0187	.0224	.0187	.0221	.0187	.0169	.0153	.2187
2	2	2	.0370	.0187	.0056	.0051	.0061	.0052	.0049	.0043	.0482
2	2	3	.0370	.0187	.0206	.0235	.0278	.0249	.0254	.0277	.2307
2	3	1	.0370	.0507	.0608	.0915	.0933	.0790	.0897	.0931	.5817
2	3	2	.0370	.0507	.0152	.0196	.0200	.0170	.0200	.0217	.2169
2	3	3	.0370	.0507	.0599	.0380	.0387	.0347	.0313	.0259	.1993
3	1	1	.0370	.0113	.0039	.0026	.0022	.0016	.0014	.0007	.0083
3	1	2	.0370	.0113	.0039	.0035	.0029	.0023	.0021	.0009	.0077
3	1	3	.0370	.0113	.0239	.0341	.0288	.0293	.0317	.0291	.2079
3	2	1	.0370	.0280	.0096	.0081	.0080	.0058	.0052	.0030	.0430
3	2	2	.0370	.0280	.0096	.0089	.0088	.0068	.0069	.0036	.0401
3	2	3	.0370	.0280	.0591	.0675	.0672	.0683	.0696	.0072	.5848
3	3	1	.0370	.0507	.0175	.0263	.0329	.0236	.0268	.0283	.1770
3	3	2	.0370	.0507	.0175	.0225	.0282	.0219	.0257	.0279	.2793
3	3	3	.0370	.0507	.1070	.0727	.0910	.0924	.0834	.0937	.7811

TABLE 9 Agricultural Time-Sequential Classification Example

			$P[\omega_k \mid \nu_j]$			
			ω_1	ω_2	ω_3	ω_4
	Spring Class	$P[\nu_j]$	Rice	Cotton	Orchard	Fallow
ν_1	Rice	.3	.9	.0	.0	.1
ν_2	Cotton	.0	--	--	--	--
ν_3	Orchard	.2	.0	.0	1.0	.0
ν_4	Fallow	.5	.1	.7	.2	.1
	$P[\omega_k]$		.32	.35	.25	.08

TABLE 10 Elevation and Aspect Class Definitions

Code	Definition	Point Count
Elevation		
Low	<1067 m	45
Middle	1068-1524 m	26
High	>1525 m	14
Aspect		
Northeast	337.6°-112.5°	26
Neutral	122.6°-157.5°; 292.6°-337.5°	25
Southwest	157.6°-292.5°	34

Figure Captions for Prior Probabilities

FIGURE 1. Landsat Band 5 image of Doggett Creek study area, Klamath National Forest, California.

FIGURE 2. Registered continuous and tri-level elevation (upper photos) and aspect (lower photos) images of Doggett Creek study area.

FIGURE 3. Classification map based on spectral data only; accuracy, 58 percent.

FIGURE 4. Classification map based on spectral data, with elevation included by varying prior probabilities. Key to map symbols is included in Figure 3. Accuracy is 71 percent.

FIGURE 5. Classification map using spectral, elevation, and aspect data. Key to map symbols is included in Figure 3. Accuracy is 77 percent.

Figure 1.

Figure 2. (upper left)

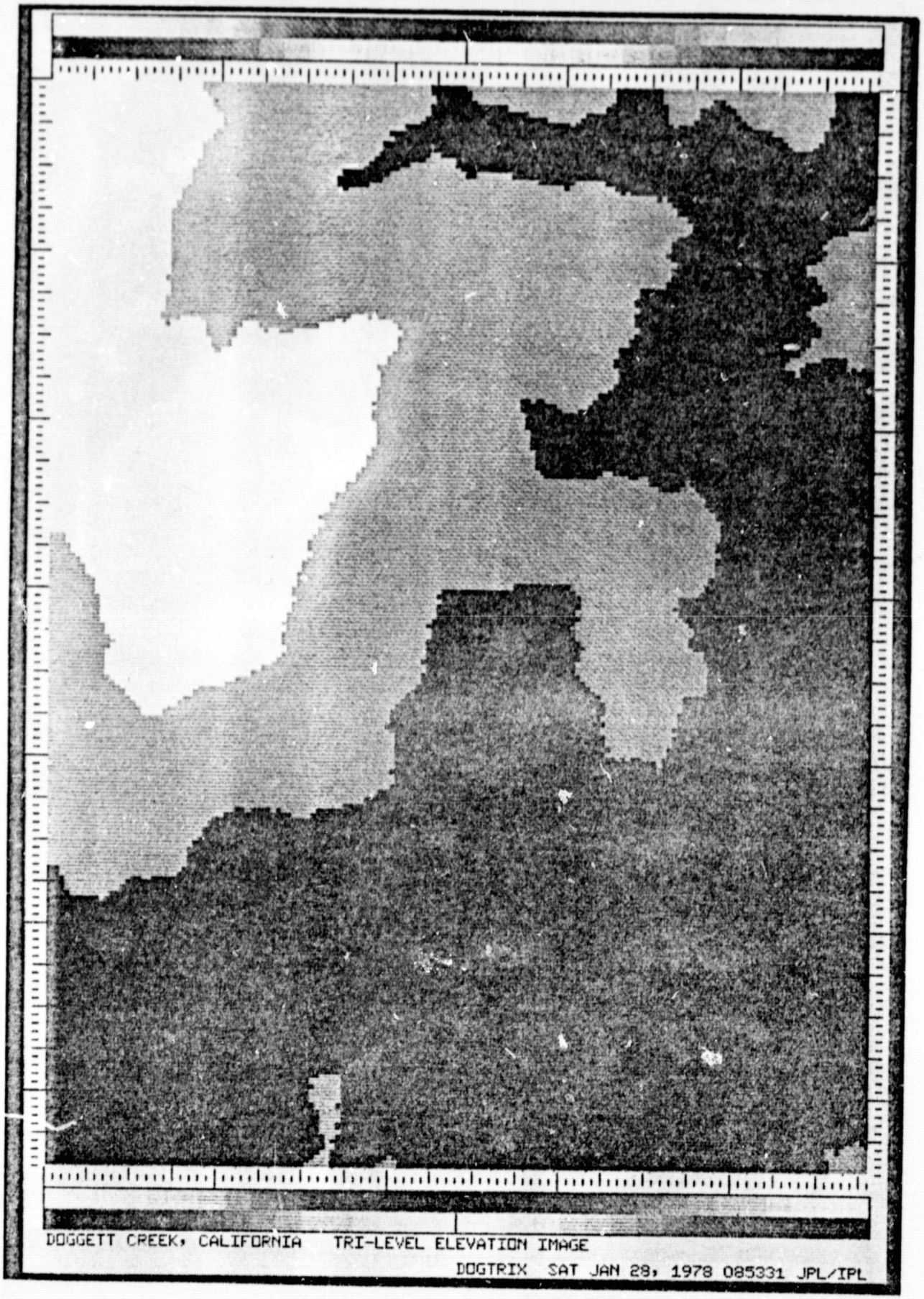

Figure 2. (upper right)

Figure 2. (lower left)

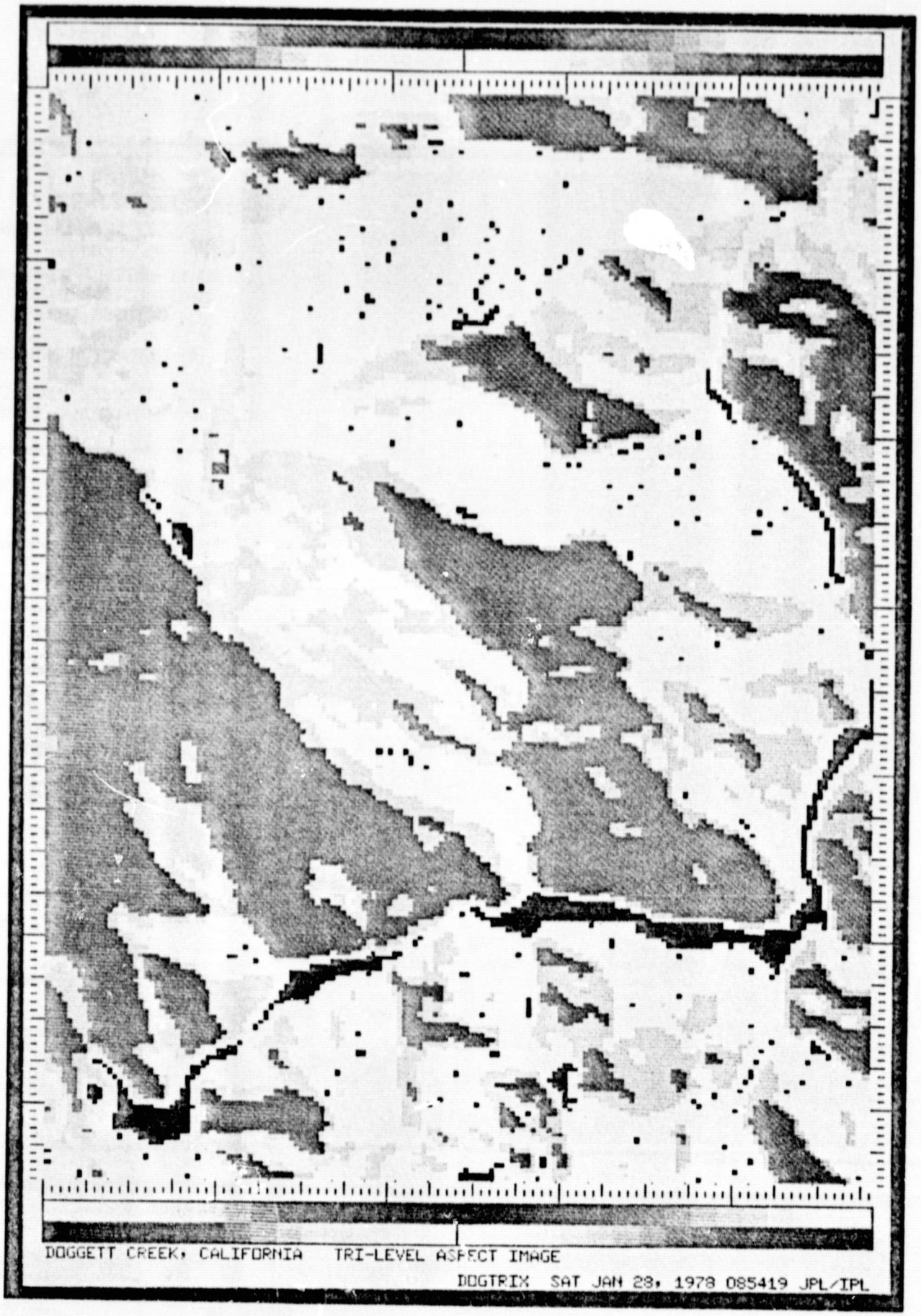

Figure 2. (lower right)

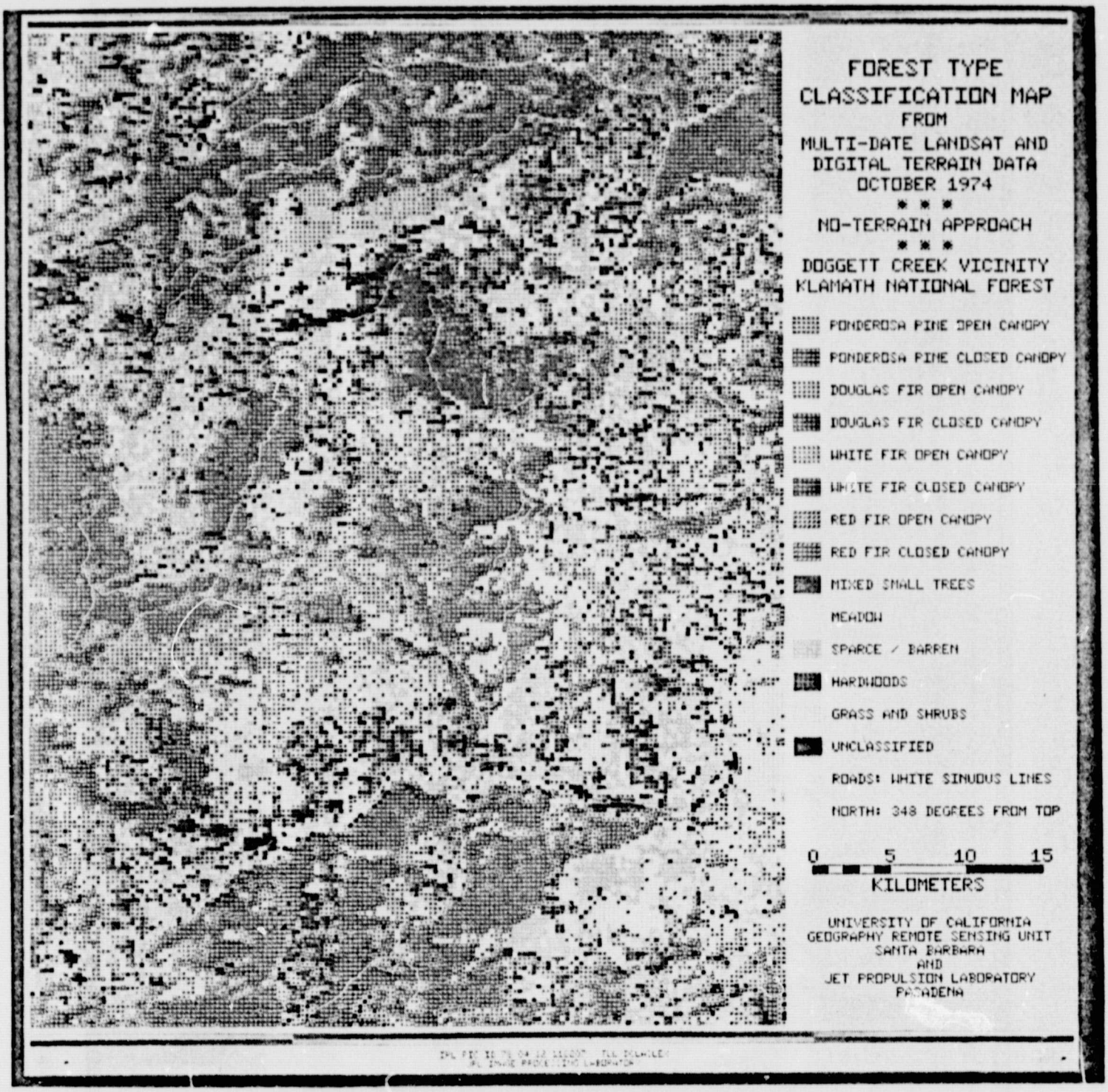

Figure 3.


IOGGETT CREEK, CALIFORNIA KLAMATH NATIONAL FOREST
PRELIMINARY 1-WAY CONDITIONAL PROB CLASS MAP
JPL PIC ID 78/03/24/134453 TLL/DONEWAYX
JPL IMAGE PROCESSING LABORATORY

Figure 4.

Figure 5.